

The **McGraw-Hill** Companies

3rd

EDITION

OPTICS



A J O Y G H A T A K

OPTICS

Third Edition

Ajoy Ghatak

*Emeritus Professor
Department of Physics
Indian Institute of Technology
Delhi*



Tata McGraw-Hill Publishing Company Limited

NEW DELHI

McGraw-Hill Offices

New Delhi New York St Louis San Francisco Auckland Bogotá Caracas
Kuala Lumpur Lisbon London Madrid Mexico City Milan Montreal
San Juan Santiago Singapore Sydney Tokyo Toronto

Information contained in this work has been obtained by Tata McGraw-Hill, from sources believed to be reliable. However, neither Tata McGraw-Hill nor its authors guarantee the accuracy or completeness of any information published herein, and neither Tata McGraw-Hill nor its authors shall be responsible for any errors, omissions, or damages arising out of use of this information. This work is published with the understanding that Tata McGraw-Hill and its authors are supplying information but are not attempting to render engineering or other professional services. If such services are required, the assistance of an appropriate professional should be sought.



Tata McGraw-Hill

Copyright © 2005, by Tata McGraw-Hill Publishing Company Limited.

Fourth reprint 2006
RYLARRBKACDL

No part of this publication may be reproduced or distributed in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise or stored in a database or retrieval system without the prior written permission of the publishers. The program listings (if any) may be entered, stored and executed in a computer system, but they may not be reproduced for publication.

This edition can be exported from India only by the publishers,
Tata McGraw-Hill Publishing Company Limited.

ISBN 0-07-058583-0

Published by the Tata McGraw-Hill Publishing Company Limited,
7 West Patel Nagar, New Delhi 110 008, typeset in Times New Roman at Tej Composers,
WZ 391, Madipur, New Delhi 110 063 and
printed at Pashupati Printers (P) Ltd. 429/16 Gali No. 1, Friends Colony,
Industrial Area, G T Road, Shahdara Delhi 110 095

Cover: De-Unique

The McGraw-Hill Companies

Contents

<i>Preface to the Third Edition</i>	v
<i>Preface to the First Edition</i>	vii
<i>Road Map to the Book</i>	viii

1. What Is Light ? **1.1**

1.1 Introduction	1.2
1.2 The Corpuscular Model	1.3
1.3 The Wave Model	1.5
1.4 The Particle Nature of Radiation	1.7
1.5 The Uncertainty Principle	1.8
1.6 The Single Slit Diffraction Experiment	1.8
1.7 The Probabilistic Interpretation of Matter Waves	1.9
1.8 An Understanding of Interference Experiments	1.11
1.9 The Polarization of a Photon	1.12
1.10 The Time–Energy Uncertainty Relation	1.14
<i>Summary</i>	1.14
<i>Problems</i>	1.15
<i>Solutions</i>	1.15
<i>References and Suggested Readings</i>	1.16

Part 1 GEOMETRICAL OPTICS

2. Fermat's Principle and its Applications **2.3**

2.1 Introduction	2.3
2.2 Laws of Reflection and Refraction from Fermat's Principle	2.4
2.3 Ray Paths in an Inhomogeneous Medium	2.8
2.4 The Ray Equation and its Solutions	2.12
2.5 Refraction of Rays at the Interface between an Isotropic Medium and an Anisotropic Medium	2.18
<i>Summary</i>	2.21
<i>Problems</i>	2.21
<i>References and Suggested Readings</i>	2.25

3. Refraction and Reflection by Spherical Surfaces **3.1**

3.1 Introduction	3.1
3.2 Refraction at a Single Spherical Surface	3.1
3.3 Reflection by a Single Spherical Surface	3.3
3.4 The Thin Lens	3.4
3.5 The Principal Foci and Focal Lengths of a Lens	3.5
3.6 The Newton Formula	3.7
3.7 Lateral Magnification	3.7
3.8 Aplanatic Points of a Sphere	3.8
3.9 The Cartesian Oval	3.10
3.10 Geometrical Proof for the Existence of Aplanatic Points	3.11

3.11 The Sine Condition	3.12
Summary	3.13
Problems	3.13
References and Suggested Readings	3.15

4. The Matrix Method in Paraxial Optics 4.1

4.1 Introduction	4.1
4.2 The Matrix Method	4.2
4.3 Unit Planes	4.7
4.4 Nodal Planes	4.8
4.5 A System of Two Thin Lenses	4.10
Summary	4.12
Problems	4.12
References and Suggested Readings	4.13

5. Aberrations 5.1

5.1 Introduction	5.1
5.2 Chromatic Aberration	5.1
5.3 Monochromatic Aberrations	5.4
Summary	5.13
Problems	5.13
References and Suggested Readings	5.14

Part 2 VIBRATIONS AND WAVES

6. Simple Harmonic Motion, Forced Vibrations and Origin of Refractive Index 6.3

6.1 Introduction	6.3
6.2 Simple Harmonic Motion	6.3
6.3 Damped Simple Harmonic Motion	6.7
6.4 Forced Vibrations	6.9
6.5 Origin of Refractive Index	6.11
6.6 Rayleigh Scattering	6.15
Summary	6.16
Problems	6.17
References and Suggested Readings	6.19

7 Fourier Series and Applications 7.1

7.1 Fourier Analysis	7.1
7.2 Transverse Vibrations of a Plucked String	7.3
7.3 Application of Fourier Series in Forced Vibrations	7.5
7.4 The Fourier Integral	7.6
Summary	7.8
Problems	7.8
References and Suggested Readings	7.9

8 Group Velocity and Pulse Dispersion 8.1

8.1 Introduction	8.1
8.2 Group Velocity	8.1
8.3 Group Velocity of a Wave Packet	8.5
8.4 Self Phase Modulation	8.12
Summary	8.14
Problems	8.15
References and Suggested Readings	8.16

9. Wave Propagation and the Wave Equation**9.1**

- [9.1 Introduction 9.1](#)
- [9.2 Sinusoidal Waves: Concept of Frequency and Wavelength 9.3](#)
- [9.3 Types of Waves 9.4](#)
- [9.4 Energy Transport in Wave Motion 9.4](#)
- [9.5 The One-Dimensional Wave Equation 9.6](#)
- [9.6 Transverse Vibrations of a Stretched String 9.6](#)
- [9.7 Longitudinal Sound Waves in a Solid 9.7](#)
- [9.8 Longitudinal Waves in a Gas 9.9](#)
- [9.9 The General Solution of the One-Dimensional Wave Equation 9.10](#)
- [Summary 9.14](#)
- [Problems 9.14](#)
- [References and Suggested Readings 9.15](#)

10. Huygens' Principle and its Applications**10.1**

- [10.1 Introduction 10.1](#)
- [10.2 Huygens' Theory 10.1](#)
- [10.3 Rectilinear Propagation 10.2](#)
- [10.4 Application of Huygens' Principle to Study Refraction and Reflection 10.4](#)
- [10.5 Huygens' Principle in Inhomogeneous Media 10.9](#)
- [Summary 10.10](#)
- [Problems 10.10](#)
- [References and Suggested Readings 10.10](#)

Part 3 INTERFERENCE**11. Superposition of Waves****11.1**

- [11.1 Introduction 11.3](#)
- [11.2 Stationary Waves on a String 11.3](#)
- [11.3 Stationary Waves on a String Whose Ends are Fixed 11.5](#)
- [11.4 Stationary Light Waves: Ives and Wiener's Experiments 11.6](#)
- [11.5 Superposition of Two Sinusoidal Waves 11.6](#)
- [11.6 The Graphical Method for Studying Superposition of Sinusoidal Waves 11.7](#)
- [11.7 The Complex Representation 11.9](#)
- [Summary 11.9](#)
- [Problems 11.9](#)
- [References and Suggested Readings 11.10](#)

12. Two Beam Interference by Division of Wavefront**12.1**

- [12.1 Introduction 12.1](#)
- [12.2 Interference Pattern Produced on the Surface of Water 12.1](#)
- [12.3 Coherence 12.5](#)
- [12.4 Interference of Light Waves 12.6](#)
- [12.5 The Interference Pattern 12.7](#)
- [12.6 The Intensity Distribution 12.8](#)
- [12.7 Fresnel's Two-Mirror Arrangement 12.14](#)
- [12.8 Fresnel Biprism 12.15](#)
- [12.9 Interference with White Light 12.15](#)
- [12.10 Displacement of Fringes 12.16](#)
- [12.11 The Lloyd's Mirror Arrangement 12.17](#)
- [12.12 Phase Change on Reflection 12.17](#)

[Summary 12.18](#)

[Problems 12.18](#)

[References and Suggested Readings 12.20](#)

13. Interference by Division of Amplitude

13.1

[13.1 Introduction 13.1](#)

[13.2 Interference by a Plane Parallel Film When Illuminated by a Plane Wave 13.1](#)

[13.3 The Cosine Law 13.3](#)

[13.4 Non-Reflecting Films 13.4](#)

[13.5 High Reflectivity by Thin Film Deposition 13.7](#)

[13.6 Reflection by a Periodic Structure 13.8](#)

[13.7 Interference by a Plane Parallel Film When Illuminated by a Point Source 13.11](#)

[13.8 Interference by a Film with Two Non-Parallel Reflecting Surfaces 13.13](#)

[13.9 Colours of Thin Films 13.16](#)

[13.10 Newton's Rings 13.16](#)

[13.11 The Michelson Interferometer 13.21](#)

[Summary 13.24](#)

[Problems 13.25](#)

[References and Suggested Readings 13.26](#)

14. Multiple Beam Interferometry

14.1

[14.1 Introduction 14.1](#)

[14.2 Multiple Reflections from a Plane Parallel Film 14.1](#)

[14.3 The Fabry-Perot Etalon 14.3](#)

[14.4 The Fabry-Perot Interferometer 14.5](#)

[14.5 Resolving Power 14.7](#)

[14.6 The Lummer-Gehrcke Plate 14.9](#)

[14.7 Interference Filters 14.10](#)

[Summary 14.11](#)

[Problems 14.11](#)

[References and Suggested Readings 14.12](#)

15. Coherence

15.1

[15.1 Introduction 15.1](#)

[15.2 The Linewidth 15.3](#)

[15.3 The Spatial Coherence 15.4](#)

[15.4 Michelson Stellar Interferometer 15.6](#)

[15.5 Optical Beats 15.7](#)

[15.6 Coherence Time and Linewidth Via Fourier Analysis 15.10](#)

[15.7 Complex Degree of Coherence and Fringe Visibility in Young's Double-Hole Experiment 15.11](#)

[15.8 Fourier Transform Spectroscopy 15.13](#)

[Summary 15.18](#)

[Problems 15.18](#)

[References and Suggested Readings 15.19](#)

Part 4 DIFFRACTION

16. Fraunhofer Diffraction

16.3

[16.1 Introduction 16.3](#)

[16.2 Single-Slit Diffraction Pattern 16.4](#)

[16.3 Diffraction by a Circular Aperture 16.8](#)

[16.4 Directionality of Laser Beams 16.10](#)

[16.5 Limit of Resolution 16.14](#)

16.6	Two-Slit Fraunhofer Diffraction Pattern	16.17
16.7	N-Slit Fraunhofer Diffraction Pattern	16.20
16.8	The Diffraction Grating	16.23
16.9	Oblique Incidence	16.26
16.10	X-Ray Diffraction	16.27
16.11	The Self-Focusing Phenomenon	16.30
16.12	Spatial Frequency Filtering	16.32
16.13	The Fourier Transforming Property of a Thin Lens	16.35
	Summary	16.38
	Problems	16.38
	References and Suggested Readings	16.40

17. Fresnel Diffraction

17.1

17.1	Introduction	17.1
17.2	Fresnel Half-Period Zones	17.2
17.3	The Zone-Plate	17.4
17.4	Fresnel Diffraction—A More Rigorous Approach	17.6
17.5	Gaussian Beam Propagation	17.8
17.6	Diffraction by a Straight Edge	17.11
17.7	Diffraction of a Plane Wave by a Long Narrow Slit and Transition to the Fraunhofer Region	17.16
17.8	Fraunhofer Diffraction by an Aperture	17.19
	Summary	17.22
	Problems	17.22
	References and Suggested Readings	17.24

18. Holography

18.1

18.1	Introduction	18.1
18.2	Theory	18.2
18.3	Requirements	18.6
18.4	Some Applications	18.7
	Summary	18.9
	Problems	18.9
	References and Suggested Readings	18.9

Part 5 ELECTROMAGNETIC CHARACTER OF LIGHT

19. Polarization and Double Refraction

19.3

19.1	Introduction	19.3
19.2	Production of Polarized Light	19.6
19.3	Malus' Law	19.8
19.4	Superposition of Two Disturbances	19.9
19.5	The Phenomenon of Double Refraction	19.12
19.6	Interference of Polarized Light: Quarter Wave Plates and Half Wave Plates	19.16
19.7	Analysis of Polarized Light	19.19
19.8	Optical Activity	19.20
19.9	Change in the Sop (State of Polarization) of a Light Beam Propagating through an Elliptic Core Single Mode Optical Fiber	19.20
19.10	Wollaston Prism	19.23
19.11	Rochon Prism	19.24
19.12	Plane Wave Propagation in Anisotropic Media	19.24
19.13	Ray Velocity and Ray Refractive Index	19.29
19.14	Jones Calculus	19.30
19.15	Faraday Rotation	19.32
19.16	Theory of Optical Activity	19.33
	Summary	19.35

[Problems](#) 19.36

[References and Suggested Readings](#) 19.38

20. Electromagnetic Waves

20.1

[20.1 Maxwell's Equations](#) 20.1

[20.2 Plane Waves in a Dielectric](#) 20.2

[20.3 The Three-Dimensional Wave Equation in a Dielectric](#) 20.4

[20.4 The Poynting Vector](#) 20.5

[20.5 Energy Density and Intensity of an Electromagnetic Wave](#) 20.8

[20.6 Radiation Pressure](#) 20.9

[20.7 The Wave Equation in a Conducting Medium](#) 20.10

[20.8 The Continuity Conditions](#) 20.11

[20.9 Physical Significance of Maxwell's Equations](#) 20.12

[Summary](#) 20.14

[Problems](#) 20.14

[References and Suggested Readings](#) 20.15

21. Reflection and Refraction of Electromagnetic Waves

21.1

[21.1 Introduction](#) 21.1

[21.2 Reflection at an Interface of Two Dielectrics](#) 21.1

[21.3 Reflection by a Conducting Medium](#) 21.13

[21.4 Reflectivity of a Dielectric Film](#) 21.14

[Summary](#) 21.16

[Problems](#) 21.16

[References and Suggested Readings](#) 21.17

Part 6 PHOTONS

22. The Particle Nature of Radiation

22.3

[22.1 Introduction](#) 22.3

[22.2 The Photoelectric Effect](#) 22.4

[22.3 The Compton Effect](#) 22.6

[22.4 The Photon Mass](#) 22.9

[Summary](#) 22.10

[Problems](#) 22.10

[References and Suggested Readings](#) 22.10

Part 7 SOURCES OF COHERENT LIGHT

23. Lasers: An Introduction

23.3

[23.1 Introduction](#) 23.3

[23.2 The Fiber Laser](#) 23.8

[23.3 The Ruby Laser](#) 23.9

[23.4 The He-Ne Laser](#) 23.10

[23.5 Optical Resonators](#) 23.11

[23.6 Einstein Coefficients and Optical Amplification](#) 23.15

[23.7 The Line-Shape Function](#) 23.20

[23.8 Typical Parameters for a Ruby Laser](#) 23.22

[23.9 Monochromaticity of the Laser Beam](#) 23.23

[23.10 Raman Amplification and Raman Laser](#) 23.24

[Summary](#) 23.27

[Problems](#) 23.28

[References and Suggested Readings](#) 23.29

Part 8 SOME CONTEMPORARY TOPICS

24. Fiber Optics	24.3
24.1 Introduction	24.3
24.2 Some Historical Remarks	24.4
24.3 Total Internal Reflection	24.6
24.4 The Optical Fiber	24.7
24.5 Why Glass Fibers?	24.8
24.6 The Coherent Bundle	24.9
24.7 The Numerical Aperture	24.10
24.8 Attenuation in Optical Fibers	24.11
24.9 Single-Mode and Multimode Fibers	24.12
24.10 Pulse Dispersion in Optical Fibers	24.14
24.11 Dispersion and Maximum Bit Rates	24.17
24.12 Waveguide Dispersion	24.18
24.13 Dispersion Compensating Fibers	24.20
24.14 Fiber-Optic Sensors	24.21
Summary	24.22
Problems	24.23
References and Suggested Readings	24.23
25. Introduction to Speckle Metrology	25.1
Summary	25.3
References and Suggested Readings	25.3
Appendix A Gamma Functions and Integrals Involving Gaussian Functions	A.1
Appendix B Diffraction of a Gaussian Beam	B.1
Index	I.1

Chapter 1

What is Light ?

For the rest of my life, I will reflect on what light is

Albert Einstein, CA 1917

All the fifty years of conscious brooding have brought me no close to the answer to the question, 'What are light quanta?' Of course today every rascal thinks he knows the answer, but he is deluding himself.

Albert Einstein, 1951

Important Milestones*

- 140 AD Greek physicist Claudius Ptolemy measured the angle of refraction in water for different angles of incidence in air and made a table of it.
- 965–1020 Ibn-al-Haitham (also known as Alhazen) carried out investigations using spherical and parabolic mirrors and was aware of spherical aberration. He also investigated the magnification produced by lenses.
- 1608 Hans Lippershey was probably the first person to construct a telescope with a converging objective lens and a diverging eye lens.
- 1609 Galileo Galilei learnt how to grind and polish his own lenses and by August 1609 he had a telescope with a magnification of around eight or nine. Newton was born in the year Galileo died (1642), and it is said that these two men laid the basis for the scientific method that was to serve us well for following three centuries.
- 1621 Although the numerical table showing the variation of angle of refraction with angle of incidence was made in 140 AD, it was only in 1621 that Willebrord Snell, a Dutch mathematician, discovered the law of refraction which is now known as Snell's law.
- 1637 Descartes derived the Snell's law; his derivation assumed corpuscular model of light which is usually attributed to Newton (Newton was only 8 years old when Descartes died so, most certainly, Descartes did not get the corpuscular model from him).
- 1657 Pierre de Fermat enunciated his principle of 'least time' and derived Snell's law of refraction and showed that if the velocity of light in the second medium is less, the ray would bend towards the normal, contrary to what predicted by the 'corpuscular theory'.
- 1665 Francisco Grimaldi was probably the first to observe the diffraction of light, i.e., the presence of light in the geometrical shadow. Later Robert Hooke also observed this phenomenon.
- 1666 Newton studied refraction as a function of colour.
- 1669 Erasmus Bartholinus discovered double refraction in calcite.
- 1672 In his first scientific paper, Newton said that white light is actually a mixture of all the colours of the rainbow.
- 1676 Olaf Roemer was the first to determine the speed of light by measuring the times at which satellites of Jupiter were eclipsed by the planet.
- 1678 Christiaan Huygens put forward the wave theory of light in a communication to the Academie des Science in Paris; he also gave the theory of double refraction in calcite discovered by Bartholinus
- 1704 Isaac Newton published his book entitled OPTICKS in which he put forward the corpuscular theory of light; in that book he also explained the formation of the rainbow.
- 1801 Thomas Young performed the famous interference experiment and enunciated the superposition principle.
- 1816 Augustin Fresnel developed the theory of diffraction using the wave theory of light.
- 1817 Using Fresnel's theory, Poisson predicted a bright spot at the center of the shadow of an opaque disc—this is usually referred to as the "Poisson spot".

Contd.

*Some of the dates have been taken from information available from the Internet.

- 1818 Fresnel and Arago carried out the experiment to demonstrate the existence of the Poisson spot validating the wave theory.
- 1819 Joseph Fraunhofer demonstrated the diffraction of light by gratings which were initially made by winding fine wires around parallel screws.
- 1823 Fraunhofer published his theory of diffraction.
- 1832 Fresnel derived the expressions for the reflection and transmission coefficients.
- 1864 James Clerk Maxwell predicted the existence of electromagnetic waves and said 'that light itself is an electromagnetic disturbance'.
- 1881 A.A. Michelson built the famous interferometer which was later called the *Michelson interferometer*. He was awarded the 1907 Nobel Prize in Physics "for his optical precision instruments and the spectroscopic and metrological investigations carried out with their aid". Michelson was America's first Nobel prize winner in science and during the presentation ceremony of the Nobel prize, the President of the Royal Swedish Academy of Sciences said, "Professor Michelson, Your interferometer has rendered it possible to obtain a non-material standard of length, possessed of a degree of accuracy never hitherto attained. By its means we are enabled to ensure that the prototype of the meter has remained unaltered in length, and to restore it with absolute infallibility, supposing it were to get lost"
- 1887 Heinrich Hertz discovered the photoelectric effect in which a metal surface irradiated by light beam would emit electrons if the frequency of the radiation was above a certain critical value.
- 1888 Hertz produced and detected electromagnetic waves of frequencies much smaller than those of light. He showed that the speed of the electromagnetic waves was the same as that of light and provided dramatic confirmation of Maxwell's electromagnetic theory.
- 1900 In order to derive the blackbody radiation formula, Planck made a drastic assumption that the oscillators can only assume discrete energies.
- 1905 In a paper entitled *On a heuristic point of view about the creation and conversion of light*, Einstein introduced the light quanta. In this paper he wrote that for explanation of phenomena like blackbody radiation, production of electrons by ultraviolet light (which is the photoelectric effect) it is necessary to assume that 'when a light ray starting from a point is propagated, the energy is not continuously distributed over an ever increasing volume, but it consists of a finite number of energy quanta, localized in space, which move without being divided and which can be absorbed or emitted only as a whole'. Einstein received the 1921 Nobel prize in Physics for his discovery of the law of photoelectric effect and *not* for his theory of relativity.
- 1905 In a paper entitled "On the electrodynamics of moving bodies", Einstein wrote *These two postulates* [of the principle of relativity] *suffice for the attainment of a simple and consistent electrodynamics of moving bodies based on Maxwell's theory for bodies at rest. The introduction of a "Light ether" will prove to be superfluous,*
- 1923 The Compton effect was discovered which could be explained by assuming the photon having momentum equal to h/λ .
- 1926 Gilbert Lewis, an American chemist, coined the word 'photon' to describe Einstein's 'localized energy quanta'.
- 1928 Dirac put forward the quantum theory of radiation.
- 1947 Dennis Gabor discovered holography.
- 1960 The first laser was fabricated by Maiman.

1.1 INTRODUCTION

As we know, the values of the mass and charge of electrons, protons, alpha-particles, etc. are known to a tremendous degree of accuracy—approximately one part in a billion! Their velocities can also be changed by the application of electric and magnetic fields. Thus, we usually tend to visualize them as tiny particles. However, they also exhibit diffraction and other effects which can only be explained if we assume them to be 'waves'. Similarly, as we shall see later in this book, there are experiments on interference, diffraction, etc., which can only be explained if we assume

a wave model of light. On the other hand, there are also such phenomena as the photoelectric effect, which can only be explained if we assume a particle model of light. Thus, the answers to questions like 'What is an electron or what is light' are very difficult. Indeed electrons, protons, neutrons, photons, alpha-particles, etc. are *neither particles nor waves*. The modern quantum theory describes them in a very abstract way which cannot be connected with everyday experience. To quote Feynman¹:

Newton thought that light was made up of particles, but then it was discovered that it behaves like a wave. Later, however (in the beginning of the

twentieth century), it was found that light did indeed sometimes behave like a particle. Historically, the electron, for example, was thought to behave like a particle, and then in many respects it behaved like a wave. So it really behaves like neither. Now we have given up. We say: 'it is like neither'.

There is one lucky break, however—electrons behave just like light. The quantum behaviour of atomic objects (electrons, protons, neutrons, photons, and so on) is the same for all, they are all 'particle-waves', or whatever you want to call them.

In this chapter, we will make a brief historical survey of the important experiments which led to models regarding the nature of light. Near the end of the chapter, we will qualitatively discuss how the wave and the particle aspects of radiation can be explained on the basis of the uncertainty principle and the probabilistic interpretation of matter waves.

1.2 THE CORPUSCULAR MODEL

The corpuscular model is perhaps the simplest model of light. According to it, a luminous body emits a stream of particles in all directions. Isaac Newton, in his book *Opticks*, wrote 'Are not the ray of light very small bodies emitted from shining substance?'. The particles are assumed to be very tiny so that when two light beams overlap, a collision between the two particles rarely occurs. Using the corpuscular model, one can explain the laws of reflection and refraction in the following manner.

The reflection law follows considering the elastic reflection of a particle by a plane surface. In order to understand refraction we consider the incidence of a particle at a plane surface ($y = 0$) as shown in Fig. 1.1; we are assuming that the motion is confined to the $x - y$ plane. The trajectory of the particle is determined by the conservation of the x -component of the momentum ($= p \sin \theta$) where θ is the angle that the direction of propagation makes with the y -axis. The conservation condition leads to the following equation:

$$p_1 \sin \theta_1 = p_2 \sin \theta_2 \quad (1)$$

where the angles θ_1 and θ_2 are defined in Fig. 1.1. Equation (1) directly gives Snell's law

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{p_2}{p_1} = \frac{v_2}{v_1} \quad (2)$$

In order to understand the explanation of Snell's law of refraction using the corpuscular model, we consider a simple experiment in which a ball moving with a certain

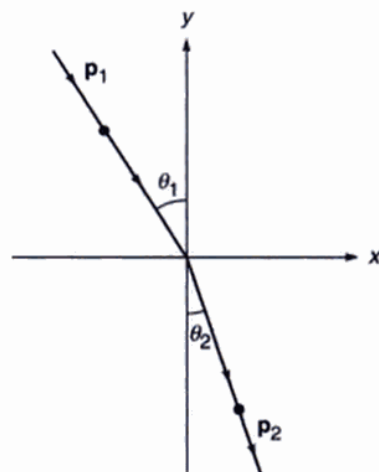


Fig. 1.1 Refraction of a corpuscle.

speed on a horizontal surface moves down to a lower horizontal surface through a slope. Two stroboscopic pictures of the motion of the ball are shown in Fig. 1.2. The component of the momentum parallel to the edge of the slope does not change; however, the component perpendicular to the edge increases in value resulting in an increased speed of the ball. Conversely, if the ball initially moves on the lower surface approaching the slope, the speed decreases as it goes up the slope; this is consistent with the reversibility of rays undergoing refraction. The slope can be approximately assumed to represent the interface between the two media.

Although the simple corpuscular model of light explains Snell's law of refraction satisfactorily, it predicts that if the ray moves towards the normal (i.e., if the refraction occurs at a denser medium) its speed would become higher which, as we shall see later, is not consistent with experimental observations. The wave theory does make the correct prediction of the ratio of velocities of waves in the two media.

It should be pointed out that contrary to popular belief, the corpuscular model of light is due to Descartes rather than to Newton. The law of refraction was discovered experimentally in 1621 by Snell (1591 – 1626). Descartes' derivation of Snell's law was published in 1637; this derivation is equivalent to the corpuscular derivation which is usually attributed to Newton. Newton (1642 – 1727) was only about eight years old when Descartes (1596 – 1650) died and the first edition of Newton's *Opticks* (in which he had discussed the corpuscular model) was published in 1704². It is probably because of the popularity Newton's *Opticks* that the corpuscular theory is usually attributed to Newton; an English translation of Descartes' original paper appears in a paper by Joyce and Joyce³. Descartes' theory remained undisputed until about 1657 when Fermat (1605 –

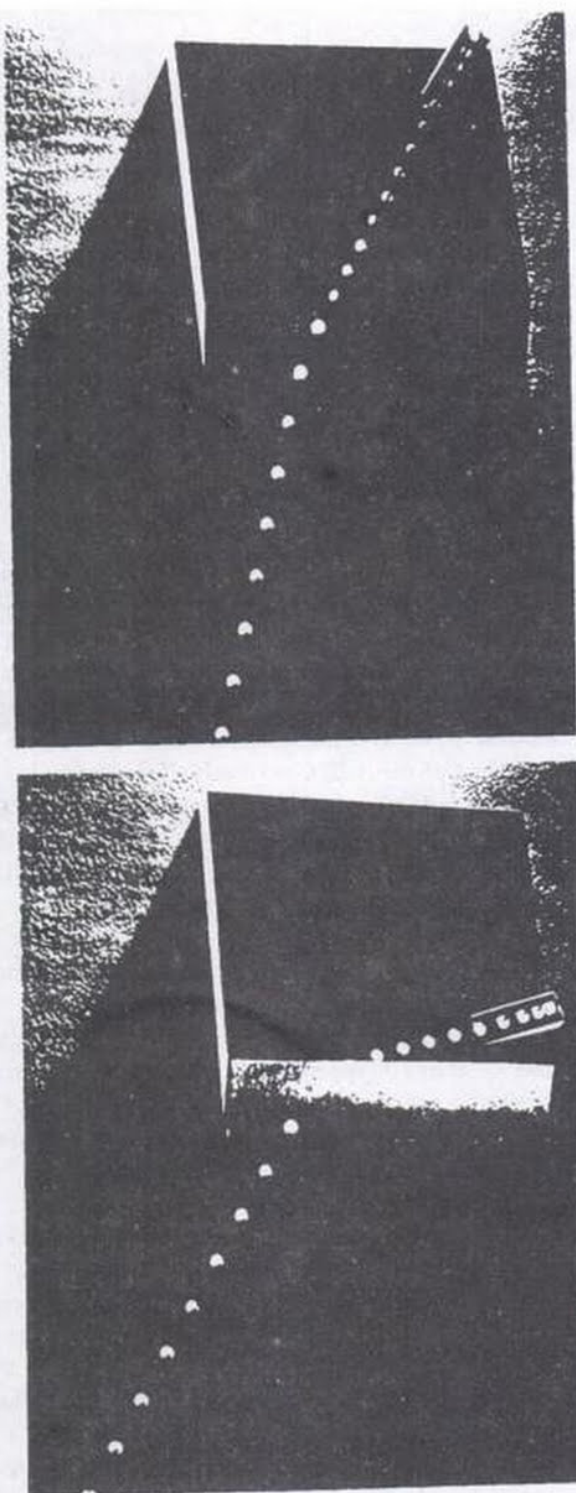


Fig. 1.2 Stroboscopic pictures of a ball moving with a certain speed on a horizontal surface moving down to a lower horizontal surface through a slope [Adapted from PSSC, Physics, D.C. Heath & CO., Boston, Mass., 1965; used with permission].

1665) enunciated the principle of least time. Using this principle, Fermat derived Snell's law and showed that if the velocity of light in the second medium is less, the ray would bend towards the normal, contrary to what predicted by the 'corpuscular theory'—see Sec. 2.2.

Finally, according to Newton, corpuscles of different sizes give rise to the sensation of different colours at the retina of the eye. He explained the prismatic spectrum by assuming that particles of different sizes refract at different angles. Commenting on Isaac Newton's *Opticks*, Einstein wrote:

.....this new edition of his [Newton's] work on Optics is nevertheless to be welcomed with warmest thanks, because it alone can afford us the enjoyment of a look at the personal activity of this unique man....In one person he [Newton] combined the experimenter, the theorist, the mechanic and, not the least, the artist in exposition. He stands before us strong, certain and alone: his joy in creation and his minute precision are evident in every word and in every figure.

Perhaps the two most important experimental facts which led to the early belief in the corpuscular model of light were: (a) the rectilinear propagation of light which results in the formation of sharp shadows, and (b) that light could propagate through vacuum. The domain of optics in which light is assumed to travel in straight lines is known as **geometrical optics** which can easily be explained on the basis of the corpuscular model of light. However, as careful experiments later showed, shadows are not perfectly dark; some light does enter the geometrical shadow which is due to the phenomenon of diffraction. This phenomenon is essentially due to the wave character of light and cannot be explained on the basis of the simple corpuscular model. As we shall see in later chapters, diffraction effects are usually difficult to observe because the wavelength associated with light waves is extremely small. We should mention here that if we are below the shade of a building then under the shade we can always read a book! The light that enters the shadow is *not* due to diffraction but due to scattering of light by air molecules. This phenomenon of scattering is also responsible for the blue color of the sky and the red color of the setting sun. If the earth did not have an atmosphere, then the shadows would have been extremely dark which is indeed the case on the surface of the moon. Since the moon does not have an atmosphere, the shadows are extremely dark and we would never be able to read a book in our own shadow! And also, on the surface of the moon, the sky appears perfectly dark (see Fig. 1.3). Once again, even on the surface of the moon, a small amount of light does enter the geometrical shadow because of diffraction.

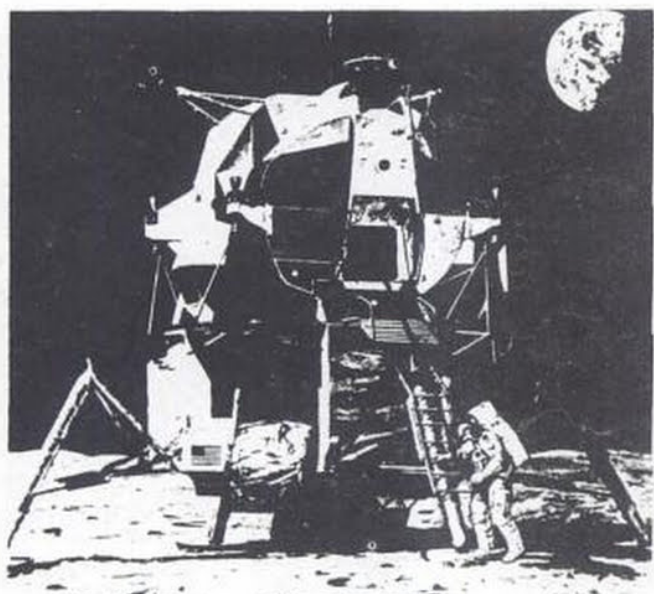


Fig. 1.3 A photograph of the man on the moon. Notice the dark sky. [Photograph courtesy: United States Information Service, New Delhi].

1.3 THE WAVE MODEL

Although the corpuscular model explains the propagation of light through free space and can be made to predict the correct forms of the laws of reflection and refraction, there came up a large number of experimental observations (like interference, diffraction and polarization) which could not be explained on the basis of the corpuscular model of light. Historically, 'Newton's rings', which are a beautiful manifestation of the wave character of light, were observed by Boyle (1627 – 1691) and by Hooke (1635 – 1703) around the middle of the seventeenth century; the rings are named after Newton because he had given an explanation of their formation, using the corpuscular model, which was later found to be quite unsatisfactory. The explanation of Newton's rings on the basis of the wave model is discussed in Chapter 12; Newton's explanation of the rings can be found in Ref. 4.

Around 1665, Francesco Grimaldi, an Italian physicist, was probably the first person to observe the phenomenon of diffraction of white light as it passed through small apertures. Grimaldi concluded that—to quote from the internet—*light is a fluid that exhibits wave-like motion*. Later, Hooke also observed this phenomenon. As will be discussed in later chapters, a satisfactory explanation of

Newton's rings and of the diffraction phenomenon can only be given if one assumes a wave model of light. This model was first propounded by Huygens in 1678⁵. Using the wave model, Huygens could explain the laws of reflection and refraction (see Chapter 10) and he could also interpret the phenomenon of double refraction (see Chapter 19) discovered in 1669 by the Danish physicist Erasmus Bartholinus (1625 – 1698). However, so compelling was Newton's authority that people around Newton had more faith in his corpuscular theory than Newton himself and no one really believed in Huygens' wave theory until 1801 when Thomas Young (1773 – 1829) performed the famous interference experiment which could only be explained on the basis of a wave model of light (see Chapter 12). In addition, at the time of Huygens, light was thought to travel in straight lines and Huygens tried to invoke unrealistic assumptions in order to explain the rectilinear propagation of light using his wave theory. This drawback was also one of the reasons for the immediate non-acceptance of the wave model. Young showed that the wavelength of light waves was about 6×10^{-5} cm. Because of the smallness of the wavelength, the diffraction effects are small, and therefore light approximately travels in straight lines. Indeed, the branch of optics in which one completely neglects the finiteness of the wavelength is called geometrical optics and a ray is defined as the path of energy propagation in the limit of $\lambda \rightarrow 0$.

In 1802, Young gave a satisfactory explanation of the formation of Newton's rings. In 1808, Malus (1775 – 1812) observed the polarization of light but he did not try to interpret this phenomenon. In 1816, Fresnel (1788 – 1827) gave a satisfactory explanation of the diffraction phenomenon by means of a wave theory and calculated the diffraction patterns produced by various types of apertures and edges. In 1816, Fresnel along with Arago (1786 – 1853) performed the famous experiment on the superposition of linearly polarized light waves, which was explained by Young, by assuming that light waves were transverse in character.

In the second quarter of the nineteenth century, the wave theory seemed to be well established and since it was thought that a wave required a medium for its propagation, the elastic ether theory was developed. Indeed, in 1832, Fresnel derived the expressions for the reflection and transmission coefficients* by using models for ether vibrations. Poisson (1781 – 1842), Navier (1785 – 1836), Cauchy (1789 – 1857) and many other physicists contributed to the development of the ether theory which also necessitated the development of the theory of elasticity.

* The derivation of the Fresnel laws on the basis of electromagnetic theory will be discussed in Chapter 21. A derivation similar to the original derivation is given in Ref. 6.

There were considerable difficulties in the explanation of the models and since we now know that ether does not exist, we will not go into the details of the various theories.

The nineteenth century also saw the development of electricity and magnetism. In 1820, Oersted (1777–1851) discovered that currents caused magnetic effects. Soon after, Ampere (1775–1836) found that two parallel wires carrying currents attract each other. Around 1830, Faraday (1791–1867) carried out experiments which showed that a varying magnetic field induces an electromotive force; similar experiments were also carried out by Henry around the same time and the law is also referred to as the Faraday–Henry law.

Soon afterwards, Maxwell (1831–1879) generalized Ampere's law by stating that a changing electric field can also produce a magnetic field. He summed up all the laws of electricity and magnetism in the form of equations which are now referred to as Maxwell's equations (see Chapter 20). From these equations, he derived a wave equation and predicted the existence of electromagnetic waves. From the wave equation so derived, he showed that the velocity of the electromagnetic waves can be calculated from experiments in which a certain quantity of electric charge is measured by two different methods. These measurements were carried out in 1856 by Kohlrausch (1809–1858) and Weber (1804–1891), and from their data, Maxwell found that the speed of the electromagnetic waves in air should be about 3.107×10^8 m/s. He found that this value was very close to the measured value of the speed of light which according to the measurement of Fizeau (1819–1896) in 1849 was 3.14858×10^8 m/s. The sole fact that the two values were very close to each other led Maxwell to propound his famous electromagnetic theory of light according to which, *light waves are electromagnetic waves*.

Associated with a light wave are changing electric and magnetic fields. The changing magnetic field produces a time and space varying electric field and the changing electric field produces a time and space varying magnetic field, and this results in the propagation of the electromagnetic wave even in free space. In 1888, Hertz (1857–1894) carried out experiments which could produce and detect electromagnetic waves of frequencies smaller than those of light. These waves were produced by discharging electrically charged plates through a spark gap. The frequency of the emitted electromagnetic waves depended on the values of the inductance and capacitance of the circuit. The electromagnetic waves could be detected by means of a detector and it was found that a signal was not received when the detector was placed parallel to the source*—see Fig. 1.4.

* This follows directly from the dipole radiation pattern—see Sec. 20.4.1; in Fig. 20.4, the dipole is oscillating along the z -axis and the electric field at a point on the y -axis is along the z -axis.

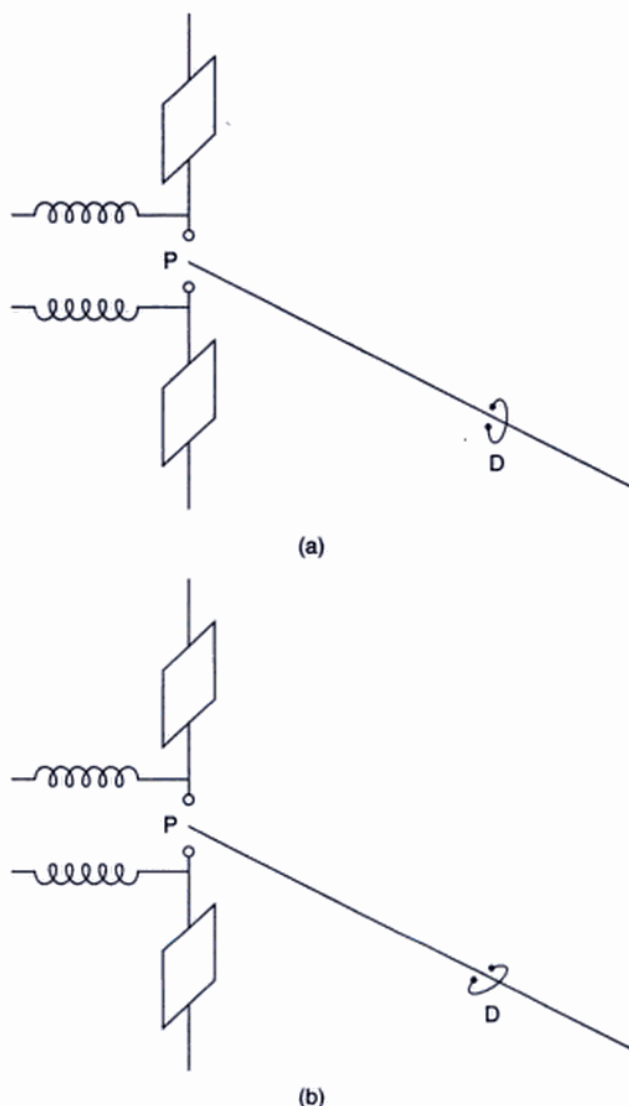


Fig. 1.4 Schematic of Hertz's experiment for generation and detection of electromagnetic waves. Sparks across the gap P produce electromagnetic waves whose frequency depends on the inductance and capacitance of the circuit. The electromagnetic waves can be detected by means of a detector D which is nothing but a short wire bent in the form of a circle with a small gap. A signal is detected if the gap in the detector is parallel to the line joining the knobs of the spark gap P as shown in (a); if the gap is at right angles as shown in (b), no signal is received.

Hertz also produced standing electromagnetic waves by getting them reflected by a metal sheet (see Figs 11.3 and

11.4). He could calculate the wavelength of the waves and knowing the frequency, he showed that the speed of the electromagnetic waves was the same as that of light. Using a collimated electromagnetic wave, and getting it reflected by a metal sheet he could demonstrate the laws of reflection. Hertz's experimental results provided a dramatic confirmation of Maxwell's electromagnetic theory. In addition, there were so many other experimental results, which were quantitatively explained by using Maxwell's theory that towards the end of the nineteenth century, physicists thought that one had finally understood what light really was.

1.4 THE PARTICLE NATURE OF RADIATION

In 1887, Hertz discovered that a metal irradiated by a light beam would emit electrons if the frequency of the radiation were above a certain critical value. This phenomenon is known as the photoelectric effect and cannot be explained by a theory based on the wave model of light; the detailed experimental findings of the photoelectric effect and the reasons why they cannot be explained on the basis of a wave model will be discussed in Section 22.2. In 1905, Einstein (1879 – 1955) interpreted the photoelectric effect by putting forward his famous photon theory according to which the energy in a light beam of frequency ν was concentrated in corpuscles of energy $h\nu$, where h represents Planck's constant. In his 1905 paper (Ref. 7—and reprinted in Ref. 8), Einstein wrote

"According to Maxwell's theory, energy is considered to be a continuous spatial function for all purely electromagnetic phenomena, hence also for light..... The wave theory of light, which operates with continuous spatial functions, has proved itself superbly in describing purely optical phenomena and will probably never be replaced by another theory. One should keep in mind, however, that optical observations, refer to time averages rather than instantaneous values; and it is quite conceivable, despite the complete confirmation of the theory of diffraction, reflection, refraction, dispersion, etc. by experiment, that the theory of light, operating with continuous spatial functions, leads to contradictions when applied to the phenomena of emission and transformation of light. Indeed, it seems to me that the observations of "blackbody radiation", photoluminescence, production of cathode rays by ultraviolet light, and other related phenomena associated with the emission or trans-

formation of light appear more readily understood if one assumes that the energy of light is discontinuously distributed in space. According to the assumption considered here, in the propagation of a light ray emitted from a point source, the energy is not distributed continuously over ever-increasing volumes of space, but consists of a finite number of energy quanta localized at points of space that move without dividing, and can be absorbed or generated only as complete units.

In this paper I wish to present the train of thought and cite the facts that led me onto this path, in the hope that the approach to be presented will prove of use to some researchers in their investigations."

We may mention here that it was only in 1926 that Gilbert Lewis, an American chemist, coined the word 'photon' to describe Einstein's 'localised energy quanta'.

Einstein received the 1921 Nobel prize in physics for his 'discovery of the law of photoelectric effect' and not for his theory of relativity. Einstein's photon theory predicted that if the frequency ν of the incident radiation was greater than the critical frequency ν_c , then the kinetic energy of the emitted electron would be $h(\nu - \nu_c)$ which was later verified by Millikan for visible light, by de Broglie for X-rays, and by Thibaud and Ellis for γ -rays. Einstein also showed that the photons, in addition to having an energy equal to $h\nu$, should have a momentum $h\nu/c$ which was verified experimentally in 1923 by Compton (1892 – 1962). This experiment is known as the Compton effect and will be discussed in detail in Sec. 22.3. Compton received the 1927 Nobel Prize in Physics 'for his discovery of the effect named after him'.

It may be of interest to mention that in 1900, Max Planck (1858 – 1947) had put forward his famous theory of blackbody radiation, the derivation of which presupposed that energy can be absorbed and emitted by an individual resonator only in 'quanta' of magnitude $h\nu$. According to Einstein

... I could nevertheless see to what kind of consequences this law (i.e., Planck's law) of temperature-radiation leads for the photoelectric effect and for other related phenomena of the transformation of radiation-energy, as well as for the specific heat of (especially) solid bodies. All my attempts however, to adapt the theoretical foundation of physics to this (new type of) knowledge failed completely. It was as if the ground had been pulled out from under one, with no firm foundation to be seen anywhere, upon which one could have built. That this insecure and contradictory foundation was sufficient to enable

*a man of Bohr's unique instinct and tact to discover the major laws of the spectral lines and of the electron shells of the atoms together with their significance for chemistry appeared to me like a miracle—and appears to me as a miracle even today. This is the highest form of musicality in the sphere of thought.**

1.5 THE UNCERTAINTY PRINCIPLE

The reconciliation of the corpuscular nature with the wave character of light (and also of the electron) has been brought about through the modern quantum theory; and perhaps the best known consequence of wave-particle duality is the uncertainty principle of Heisenberg which can be stated as follows:

If the x -coordinate of the position of a particle is known to an accuracy Δx , then the x -component of the momentum cannot be determined to an accuracy better than $\Delta p_x \approx h/\Delta x$, where h is the Planck's constant.

Alternatively, one can say that if Δx and Δp_x represent the accuracies with which the x -coordinate of the position and the x -component of the momentum can be determined, then the following inequality must be satisfied

$$\Delta x \Delta p_x \geq h \quad (3)$$

We do not feel the effect of this inequality in our everyday experience because of the smallness of the value of Planck's constant ($\approx 6.6 \times 10^{-27}$ erg sec). For example, for a tiny particle of mass 10^{-6} gm, if the position is determined within an accuracy of about 10^{-6} cm, then according to the uncertainty principle, its velocity cannot be determined within an accuracy better than $\Delta v \approx 6 \times 10^{-16}$ cm/sec. This value is much smaller than the accuracies with which one can determine the velocity of the particle. For a particle of a greater mass, Δv will be even smaller. Indeed, had the value of Planck's constant been much larger, the world would have been totally different. In a beautifully written book, Gamow⁹ has discussed what our world would be like if the effect of the uncertainty principle were perceivable by our senses.

1.6 THE SINGLE SLIT DIFFRACTION EXPERIMENT

We will now show how the diffraction of a light beam (or an electron beam) can be explained on the basis of the

corpuscular nature of radiation and the uncertainty principle. Consider a long narrow slit of width b as shown in Fig. 1.5. Now, one can always assume the distance between the source and the slit to be large enough so that p_x can be assumed to have an arbitrarily small value. For example, for the source at a distance d , the maximum value of p_x of the photons approaching the slit will be

$$p \frac{b}{d} = \frac{h\nu}{c} \cdot \frac{b}{d}$$

which can be made arbitrarily small by choosing a large enough value of d . Thus we may assume the light source to be sufficiently far away from the slit so that the photons approaching the slit can be assumed to have momentum only in the y -direction. Now, according to the particle model of radiation, the number of particles reaching the point P (which lies in the geometrical shadow) will be extremely small; further, if we decrease the width of the slit, the intensity should decrease, which is quite contrary to the experimental results because we know that the beam undergoes diffraction and the intensity at a point like P would normally increase if the width of the slit is made smaller. Thus, the classical corpuscular model is quite incapable of explaining the phenomenon of diffraction. However, if we use the uncertainty principle in conjunction with the corpuscular model, the diffraction phenomenon can be explained in the following manner: When a photon (or an electron) passes through the slit, one can say that

$$\Delta x \approx b$$

which implies that we can specify the position of the photon to an accuracy b . If we now use the uncertainty principle, we would have

$$\Delta p_x = \frac{h}{b} \quad (4)$$

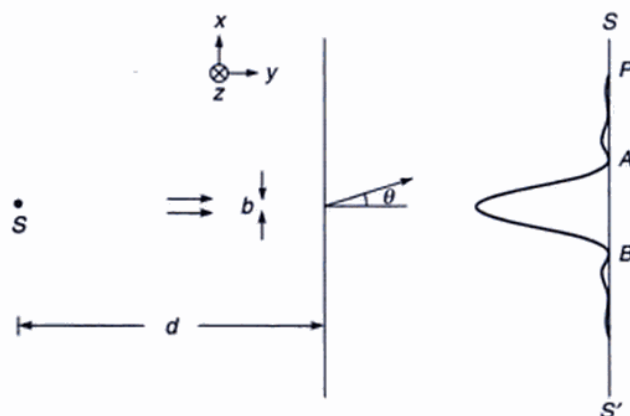


Fig. 1.5 Diffraction of a photon (or an electron) beam by a narrow slit of width b .

*Quoted from the autobiographical notes by Einstein in *Albert Einstein: Philosopher, Scientist*, edited by P. A. Schilpp, Tudor Publishing Co., New York, 1951.

i.e., just by making the photon pass through a slit of width b , the slit imparts a momentum in the x -direction which is $\approx h/b$. It may be pointed out that *before* the photon entered the slit, p_x (and hence Δp_x) can be made arbitrarily small by putting the source sufficiently far away. Thus we may write $\Delta p_x \approx 0$. It would however be wrong to say that by making the photon pass through the slit, $\Delta p_x \Delta x$ is zero; this is because of the fact that $\Delta p_x = 0$ *before* the photon entered the slit. After the photon has entered the slit, it is confined within a distance b in the x direction and hence $\Delta p_x \approx h/b$. Further, since before the photon entered the slit $p_x \approx 0$, we will therefore have

$$|p_x| \approx \Delta p_x \approx \frac{h}{b}$$

But $p_x = p \sin \theta$, where θ is the angle that the photon coming out of the slit makes with the y -axis (see Fig. 1.5). Thus

$$p \sin \theta \approx \frac{h}{b}$$

or

$$\sin \theta \approx \frac{h}{pb} \quad (6)$$

The above equation predicts that the possibility of a photon traveling at an angle θ with the y -direction is inversely proportional to the width of the slit; i.e., smaller the value of b , greater is the value of θ and greater is the possibility of the photon to reach deep inside the geometrical shadow. This is indeed the diffraction phenomenon. Now, the momentum of a photon is given by

$$p = \frac{h}{\lambda} \quad (7)$$

Thus Eq.(6) becomes

$$\sin \theta \approx \frac{\lambda}{b} \quad (8)$$

which is the familiar diffraction theory result as will be discussed in Sec. 16.3. We can therefore say that the wave-particle duality is a consequence of the uncertainty principle and the uncertainty principle is a consequence of the wave-particle duality. To quote Max Born¹⁰

Physicists of today have learnt that not every question about the motion of an electron or a photon can be answered, but only those questions which are compatible with the uncertainty principle.

Returning to Eq. (7), we may mention that de Broglie had suggested that the equation $\lambda = h/p$ is not only valid for photons but is also valid for all particles like electrons, protons, neutrons, etc. Indeed, the de Broglie relation has been verified by studying the diffraction patterns produced when electrons, neutrons, etc., pass through a single crystal; the patterns can be analyzed in a manner similar to X-ray diffractions (see Chapter 16). In Fig. 1.6, we show the

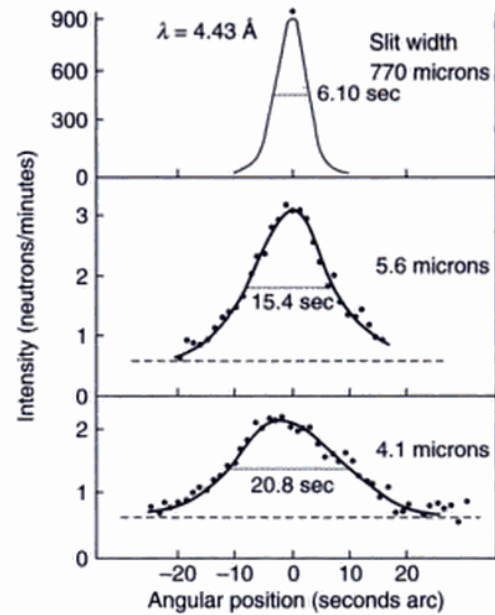


Fig. 1.6 Angular broadening of a neutron beam by small slits [After Ref. 11].

experimental data of Shull¹¹ who studied the Fraunhofer diffraction of neutrons by a single slit and his experimental results agree with the intensity distribution as predicted by the wave theory with λ given by Eq.(7).

1.7 THE PROBABILISTIC INTERPRETATION OF MATTER WAVES

In the previous section we have seen that if a photon passes through a slit of width b , then the momentum imparted in the x -direction (which is along the width of the slit) is $\approx h/b$. The question arises whether we can predict the trajectory of an individual photon. The answer is *no*. We cannot say where an individual photon will land up on the screen; we can only predict the probabilities of arrival of the photon in a certain region of the screen. We may, for example, say that the probability for the arrival of the photon in the region lying between the points A and B (see Fig. 1.5) is 0.85. This would imply that if the experiment was carried out with a large number of photons, about 85% of them would land up in the region AB; but the fate of an individual photon can never be predicted. This is in sharp contrast to Newtonian mechanics where the trajectories are always predetermined. It should be mentioned that if we place a light detector on the screen, then it will always record one photon or none and never half a photon. This essentially implies the corpuscular nature of the radiation. However, the probability distribution is the same as predicted by the wave theory and therefore if one performs an experiment with a large number of photons (as is indeed

the case in most experiments) the intensity distribution recorded on the screen is the same as predicted by the wave theory.

In order to explicitly show that diffraction is not a many photon phenomenon, Taylor in 1909 carried out a beautiful experiment which consisted of a box with a small lamp which casts the shadow of a needle on a photographic plate (see Fig. 1.7). The intensity of light was so weak that between the slit and the photographic plate, it was almost impossible to find two photons (see Example 1.1). In fact, to get a good fringe pattern, Taylor made an exposure lasting several months. The diffraction pattern obtained on the photographic plate was the same as predicted by the wave theory.

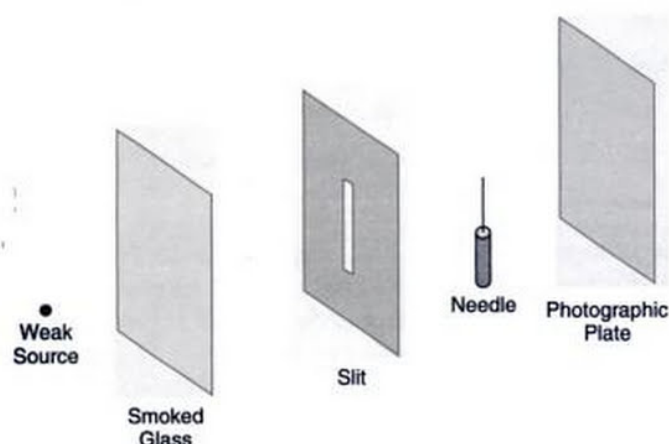


Fig. 1.7 Schematic diagram of the experimental arrangement of Taylor to study the diffraction pattern produced by a weak source. The whole apparatus was placed inside a box.

The corpuscular nature of radiation and the fact that one cannot predict the trajectory of an individual photon can be seen from Fig. 1.8, which consists of a series of photographs showing the quality of pictures obtainable from various number of photons¹². The photograph clearly shows that the picture is built up by the arrival of concentrated packets of energy and the point at which a particular photon will arrive is entirely a matter of chance. The figure also shows that the photograph is featureless when a small number of photons are involved and as the number of photons reaching the photographic plate increases, the intensity distribution becomes the same as would be predicted by the wave theory. To quote Feynman¹:

...it would be impossible to predict what would happen. We can only predict the odds! This would mean, if it were true, that physics has given up on the problem of trying to predict exactly what will

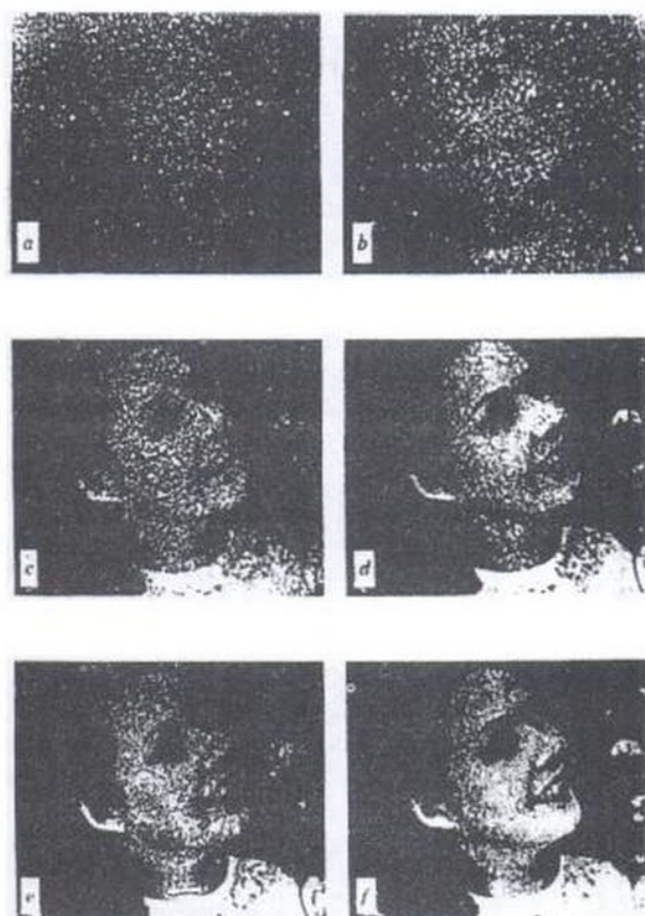


Fig. 1.8 Photographs showing the quality of a picture obtainable from various numbers of photons: (a), (b), (c), (d), (e) and (f) correspond to 3×10^3 photons, 1.2×10^4 photons, 9.3×10^4 photons, 7.6×10^5 photons, 3.6×10^6 photons and 2.8×10^7 photons respectively. [From Ref. 12; reprinted with permission].

happen in a definite circumstance. Yes! Physics has given up. We do not know how to predict what would happen in a given circumstance, and we believe now that it is impossible—that the only thing that can be predicted is the probability of different events. It must be recognized that this is a retrenchment in our earlier idea of understanding nature. It may be a backward step, but no one has seen a way to avoid it.

A somewhat similar situation arises in radioactivity. Consider a radioactive nucleus having a half-life of say 1 hour. If we start with 1,000 such nuclei, then on an average 500 of them would undergo radioactive decay in 1 hour and about 250 of them in the next 1 hour and so on. Thus, although to start with, all nuclei are identical, some nuclei

would decay in the very first minute and some nuclei can survive for hours without undergoing radioactive decay. Thus, one can never predict as to which nucleus will undergo decay in a specified period; one can only predict the probability of its undergoing decay in a certain interval of time. This is indeed a beautiful manifestation of quantum mechanics. To quote Feynman again:

A philosopher once said it is necessary for the very existence of science that the same conditions always produced the same results. Well they don't!

1.8 AN UNDERSTANDING OF INTERFERENCE EXPERIMENTS

Let us consider the interference experiment involving the Michelson interferometer in which a light beam is partially reflected by a beam splitter and the resulting beams are made to interfere (see Fig. 1.9); the interference pattern produced by a Michelson interferometer is discussed in great detail in Chapter 13. According to Dirac¹³:

... we describe the photon as going partly into each of the two components into which the incident beam is split. The photon is then, as we may say, in a translational state given by the superposition of the two translational states associated with the two components.... For a photon to be in a definite translational state it need not be associated with one single beam of light, but may be associated with two or more beams of light, which are the components into which one original beam has been split. In the accurate mathematical theory each translational state is associated with one of the wave functions of ordinary wave optics, which may describe either a single beam or two or more beams into which one original beam has been split.

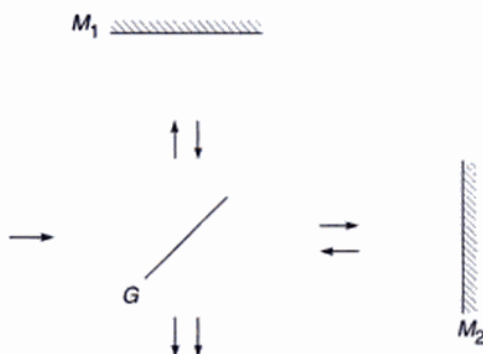


Fig. 1.9 Schematic of the set-up of the Michelson interferometer. G represents a beam splitter, M_1 and M_2 represent plane mirrors.

These translational states can be superposed in a manner similar to the one employed while considering the interference of two beams. Thus, each photon goes partly into each of the two components and interferes only with itself. If we try to determine the fate of a single photon by measuring the energy in one of the components then Dirac argues:

The result of such a determination must be either a whole photon or nothing at all. Thus the photon must change suddenly from being partly in one beam and partly in the other to be entirely in one of the beams. This sudden change is due to the disturbance in the translational state of the photon which the observation necessarily makes. It is impossible to predict in which of the two beams the photon will be found. Only the probability of either result can be calculated Our description of the photon allows us to infer that, after such an energy measurement, it would not be possible to bring about any interference effects between the two components. So long as the photon is partly in one beam and partly in the other, interference can occur when the two beams are superposed, but this probability disappears when the photon is forced entirely into one of the beams by an observation.

In a similar manner, we may consider the two-hole interference experiment similar to that performed by Young (see Sections 12.5 and 12.6). The experimental arrangement is shown in Fig. 1.10 where a weak light source S_0 illuminates the hole S and the light emerging from the holes S_1 and S_2 produces the interference pattern on the screen PP' . The intensity is assumed to be so weak that in the region between the planes AB and PP' there is almost never more than one photon (see Example 1.1). Individual photons are also counted by a detector on the screen PP' and one finds

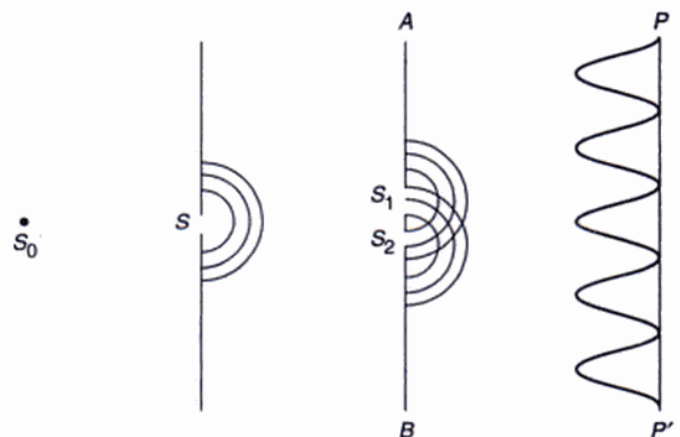


Fig. 1.10 Young's double hole experimental arrangement for obtaining the interference pattern. S_0 represents a point source.

that the intensity distribution has a \cos^2 pattern similar to that shown in Fig. 12.9 and 12.10. The corpuscular nature of the radiation is evident from its detection in the form of single photons and never a fraction of a photon. The appearance of the interference pattern is because of the fact that a photon interferes with itself. The quantum theory tells us that a photon partially passes through the hole S_1 and partially through S_2 . This is not the splitting of the photon into two halves but only implies that if we wish to find out through which hole the photon passed, then half the time it will be found to have passed through the hole S_1 and half the time through S_2 . As in the case of the Michelson interferometer, the photon is in a state which is a superposition of two states, one corresponding to the wave emanating from hole S_1 and the other to the one emanating from hole S_2 . The superposed state will give rise to an intensity distribution similar to that obtained by considering the superposition of two waves. It may be noted that if we had employed a device (like a microscope) which would have determined which hole the photon had passed through, then the interference pattern on the screen would have been washed out. This is a consequence of the fact that a measurement always disturbs the system. This is beautifully discussed in Ref. 1. Thus we may say that the photons would arrive as packets of energy but the probability distribution (on the screen) will be proportional to the intensity distribution predicted by using a wave model.

In a recent paper, Tonomura and his co-workers¹⁴ have demonstrated the single electron build-up of an interference pattern. Their results are shown in Fig. 1.11. It may be seen that when there are very few electrons they arrive randomly; however, when a large number of electrons are involved, one obtains an intensity distribution similar to the one predicted by wave theory.

1.9 THE POLARIZATION OF A PHOTON

Let us consider the incidence of a plane electromagnetic wave on a polaroid whose pass axis is along the y -direction (see Fig. 1.12); obviously, the electric vector of the transmitted wave would be along the y -direction (see Chapter 19). Thus, if the electric-vector associated with the incident wave oscillates along the x -axis, the wave will be absorbed by the polaroid. On the other hand, if the electric vector oscillates along the y -axis, it will just pass through the polaroid. Further, if the electric-vector makes an angle θ with the pass axis, then the intensity of the transmitted beam will be $I_0 \cos^2 \theta$, where I_0 represents the intensity of the incident beam (this is known as Malus' law which will be discussed in Sec. 19.3).



Fig. 1.11 Build-up of the electron interference pattern. Number of electrons in (a), (b), (c), (d) and (e) are 10, 100, 3000, 20000 and 70000 respectively. [Adapted from Ref. 14].

In the photon theory also one can associate a certain state of polarization with every photon. One can argue that if the electric-vector associated with the photon is along the y - (or the x -) axis, then the photon will pass through or get absorbed by the polaroid. The question now arises as to

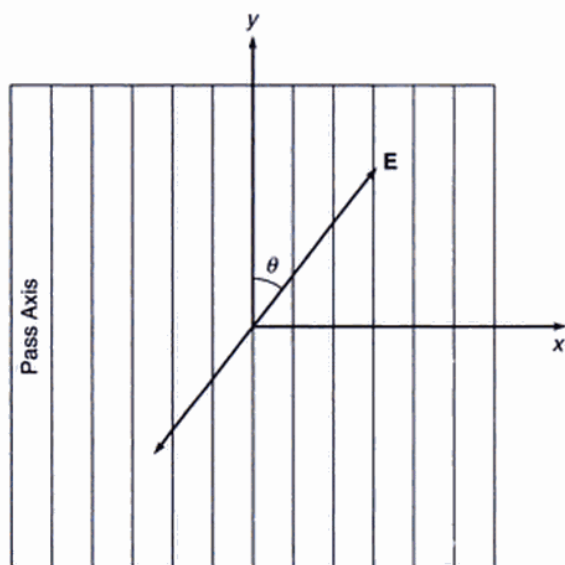


Fig. 1.12 The incidence on a polaroid of a linearly polarized light beam whose electric-vector makes an angle θ with the y-axis; the pass axis of the polaroid is along the y-axis.

what will happen to a single photon if the electric-vector makes an angle θ with the pass axis. The answer is that the probability for the photon to pass through the polaroid is $\cos^2 \theta$ and if the experiment is conducted with N photons (and if N is very large) then about $N \cos^2 \theta$ photons will pass through; one cannot predict the fate of an individual photon.

Example 1.1 Let a source (with $\lambda = 5 \times 10^{-5}$ cm) of power 1 W be used in the experimental arrangement shown in Fig. 1.10.

- Calculate the number of photons emitted by the source per second.
- Assume the radii of the holes S , S_1 and S_2 to be 0.02 cm and $S_0S = SS_1 = SS_2 = 100$ cm and the distance between the planes AB and PP' to be also 100 cm. Show that in the region between the planes AB and PP' one can almost never find two photons.

Solution:

- The energy of each photon will be

$$h\nu = \frac{hc}{\lambda} = \frac{6.6 \times 10^{-34} (\text{J} \cdot \text{s}) \times 3 \times 10^8 (\text{m/s})}{5 \times 10^{-7} (\text{m})}$$

$$\approx 4 \times 10^{-19} \text{ J}$$

Thus the number of photons emitted per second will be

$$\frac{1 \text{ W}}{4 \times 10^{-19} \text{ J}} = 2.5 \times 10^{18}$$

- The number of photons passing through the hole S will approximately be

$$\frac{2.5 \times 10^{18} \times \pi \times (0.02)^2}{4 \times \pi \times (100)^2} = 2.5 \times 10^{10} \text{ per second}$$

Similarly, the number of photons passing through either S_1 or S_2 will approximately be

$$\frac{2.5 \times 10^{10} \times 2 \times \pi \times (0.02)^2}{2 \pi \times (100)^2} = 1000 \text{ per second}$$

where we have assumed that after passing through S , the photons are evenly distributed in the hemisphere. This is strictly not correct because the diffraction pattern is actually an Airy pattern (see Chapter 16)—nevertheless, the above calculations are qualitatively correct. The distance between the planes AB and PP' is 100 cm which will be traversed by a photon in time $\approx 3 \times 10^{-9}$ s. Thus, approximately every thousandth of a second a photon enters the region and the space is traversed much before the second photon enters. Therefore, in the region between AB and PP' one will (almost) never find two photons. This is somewhat similar to the case when, on an average, 100 persons pass through a room in one year and the time that each person takes to cross the room is ≈ 1 s, thus it will be highly improbable to have two persons simultaneously in the room.

Example 1.2 In this example we will use the uncertainty principle to determine the size of the hydrogen atom. Although this example is not directly related to optics, it demonstrates the far-reaching consequences of the uncertainty principle. The analysis is adapted from Ref. 1.

We consider the hydrogen atom which consists of a proton and an electron. Since the proton is very much heavier than the electron, we consider only the motion of the electron. Let the electron be confined to a region of linear dimension $\approx a$. Thus according to the uncertainty principle:

$$p \approx \Delta p \approx \hbar/a \quad (9)$$

where $\hbar = h/2\pi$; the reason for using \hbar rather than h will be mentioned later. The kinetic energy of the electron will be given by

$$K.E. = \frac{p^2}{2m} \approx \frac{\hbar^2}{2ma^2} \quad (10)$$

Now, there exists an electrostatic attraction between the two particles; the corresponding potential energy being given by

$$P.E. = -\frac{q^2}{4\pi\epsilon_0 a} \quad (11)$$

where q ($\approx 1.6 \times 10^{-19}$ C) represents the magnitude of the charge of the electron and ϵ_0 ($\approx 8.854 \times 10^{-12}$ CN⁻² m⁻²) represents the permittivity of free space. Thus the total energy is given by

$$E = K.E. + P.E.$$

$$\approx \frac{\hbar^2}{2ma^2} - \frac{q^2}{4\pi\epsilon_0 a} \quad (12)$$

The system would settle to a state of lowest energy; thus we must set dE/da equal to zero:

$$0 = \frac{dE}{da} = -\frac{\hbar^2}{ma^3} + \frac{q^2}{4\pi\epsilon_0 a^2}$$

implying

$$a = a_0 = \frac{\hbar^2}{m \left(\frac{q^2}{4\pi\epsilon_0} \right)} \quad (13)$$

If we substitute the values of \hbar ($\approx 1.055 \times 10^{-34}$ Js), m ($\approx 9.11 \times 10^{-31}$ kg), ϵ_0 and q we would obtain

$$a = a_0 = 0.53 \times 10^{-10} \text{ m} = 0.53 \text{ \AA} \quad (14)$$

Thus we get the remarkable result that the size of the hydrogen atom is a direct consequence of the uncertainty principle. To quote Feynman:

So we now understand why we do not fall through the floor.... In order to squash the atoms close together, the electrons would be confined to smaller space and by the uncertainty principle, their momenta would have to be higher on the average, and that means high energy; the resistance to atomic compression is a quantum mechanical effect...

We next substitute the value of a from Eq. (13) in Eq. (12) to obtain

$$E = \frac{\hbar^2}{2m} \left(\frac{m}{\hbar^2} \frac{q^2}{4\pi\epsilon_0} \right)^2 - \frac{q^2}{4\pi\epsilon_0} \left(\frac{m}{\hbar^2} \frac{q^2}{4\pi\epsilon_0} \right)$$

$$= -\frac{m}{2\hbar^2} \left(\frac{q^2}{4\pi\epsilon_0} \right)^2 \quad (15)$$

Substituting the values of \hbar , m , q and ϵ_0 , we get

$$E \approx -2.17 \times 10^{-19} \text{ J}$$

$$\approx -13.6 \text{ eV} \quad (16)$$

which is nothing but the ground state energy of the hydrogen atom. Thus, that the ionization energy of hydrogen atom is ≈ 13.6 eV follows from the uncertainty principle.

We may point out that the uncertainty principle can be used to give only an order of magnitude of the size of the hydrogen atom or its ionization energy; we had intentionally chosen the constants in such a way that the ground state energy comes out to be correct. It is for this reason that we had chosen \hbar instead of h in Eqs. (9) and (10).

1.10 THE TIME-ENERGY UNCERTAINTY RELATION

When an atom makes a transition from an energy state E_2 to an energy state E_1 ($E_2 > E_1$) a photon of frequency

$$\nu = \frac{E_2 - E_1}{h}$$

is emitted. This emission is essentially a pulse of a duration $\approx 10^{-9}$ s; this duration is usually denoted by τ . This leads to a frequency width $\Delta\nu$ and as will be shown in Chapter 15,

$$\tau \Delta\nu \geq 1 \quad (17)$$

Multiplying both sides by h , we get the time-energy uncertainty principle:

$$\Delta t \Delta E \geq h \quad (18)$$

which can be interpreted as follows: if Δt represents the uncertainty in the time at which a time-dependent process takes place, then the uncertainty ΔE in the energy of this process will be $\geq h/\Delta t$. Assuming $\Delta t = 10^{-9}$ s,

$$\Delta E = \frac{h}{\Delta t} \approx 4 \times 10^{-6} \text{ eV}$$

which is known as the linewidth of the state (see Fig. 23.22).

SUMMARY

- The corpuscular model of light is due to Descartes rather than to Newton. The law of refraction was discovered experimentally in 1621 by Snell. In 1637, using corpuscular model, Descartes derived Snell's law of refraction.
- The wave model of light was first propounded by Huygens in 1678. Using the wave model, Huygens could explain the laws of reflection and refraction and he could also interpret the phenomenon of double refraction.
- Around the middle of the nineteenth century, Maxwell generalized Ampere's law by stating that a changing electric field can also produce a magnetic field. He summed up all the laws of electricity and magnetism in the form of equations which are now

referred to as Maxwell's equations. From these equations, he derived a wave equation and predicted the existence of electromagnetic waves and showed that the speed of the electromagnetic waves in air should be about 3.107×10^8 m/s which was very close to the measured value of the speed of light. The sole fact that the two values were very close to each other led Maxwell to propound his famous electromagnetic theory of light according to which, *light waves are electromagnetic waves*.

- In 1905, Einstein interpreted the photoelectric effect by putting forward his famous photon theory according to which the energy in a light beam of frequency ν was concentrated in corpuscles of energy $h\nu$, where h represents Planck's constant.
- The consequence of wave-particle duality is the uncertainty principle of Heisenberg according to which *if the x -coordinate of the position of a particle is known to an accuracy Δx , then the x -component of the momentum cannot be determined to an accuracy better than $\Delta p_x \approx h/\Delta x$, where h is the Planck's constant.*
- The classical corpuscular model is quite incapable of explaining the diffraction of light by a single slit. However, if we use the uncertainty principle in conjunction with the corpuscular model, the diffraction phenomenon can be explained.
- In the Young's double hole interference pattern, the corpuscular nature of the radiation is evident from its detection in the form of single photons and never a fraction of a photon. The appearance of the interference pattern is because of the fact that a photon interferes with itself. The quantum theory tells us that a photon partially passes through the two holes. This is not the splitting of the photon into two halves but only implies that the photon is in a state which is a superposition of two states, one corresponding to the wave emanating from the first hole and the other to the one emanating from the second hole. The superposed state will give rise to an intensity distribution similar to that obtained by considering the superposition of two waves.

PROBLEMS

- 1.1 An electron of energy 200 eV is passed through a circular hole of radius 10^{-4} cm. What is the uncertainty introduced in the momentum and also in the angle of emergence?
[Ans: $\Delta p \sim 5 \times 10^{-24}$ g cm/s; $\Delta\theta \approx 6 \times 10^{-6}$ radians]
- 1.2 In continuation of the previous problem, what would be the corresponding uncertainty for a 0.1 g lead ball

thrown with a velocity 10^3 cm/sec through a hole 1 cm in radius?

[Ans: $\Delta\theta \approx 5 \times 10^{-30}$ radians]

- 1.3 A photon of wavelength 6000\AA is passed through a slit of width 0.2 mm
- (a) Calculate the uncertainty introduced in the angle of emergence.
 - (b) The first minimum in the single slit diffraction pattern occurs at $\sin^{-1}(\lambda/b)$ where b is the width of the slit. Calculate this angle and compare with the angle obtained in part (a).
- 1.4 A 50 W bulb radiates light of wavelength $0.6 \mu\text{m}$. Calculate the number of photons emitted per second.
[Ans: $\approx 1.5 \times 10^{20}$ photons/s]
- 1.5 Calculate the uncertainty in the momentum of a proton which is confined to a nucleus of radius equal to 10^{-13} cm. From this result, estimate the kinetic energy of the proton inside the nucleus. What would be the kinetic energy for an electron if it had to be confined within a similar nucleus?
- 1.6 The lifetime of the 2P state of the hydrogen atom is about 1.6×10^{-9} s. Use the time energy uncertainty relation to calculate the frequency width $\Delta\nu$.
[Ans: $\approx 6 \times 10^8 \text{ s}^{-1}$]
- 1.7 A 1 W laser beam (of diameter 2 cm) falls normally on two circular holes each of diameter 0.05 cm as shown in Fig. 1.13. Calculate the average number of photons that will be found between the planes AB and PQ. Assume $\lambda = 6 \times 10^{-5}$ cm and the distance between the planes AB and PQ to be 30 cm.
[Ans: $\approx 4 \times 10^6$ photons]

SOLUTIONS

- 1.5 The proton is confined within a sphere of radius $r_0 \approx 10^{-13}$ cm. Thus the uncertainty in the momentum must be at least of the order of \hbar/r_0 , or

$$p \sim \frac{\hbar}{r_0}$$

Therefore, the kinetic energy of the proton will be given by

$$E = \frac{p^2}{2m_p} \sim \frac{\hbar^2}{2m_p r_0^2}$$

where m_p is the mass of the proton. On substitution, we get

$$E \sim \frac{(1.05 \times 10^{-27} \text{ erg-sec})^2}{2 \times 1.67 \times 10^{-24} \text{ g} \times (10^{-13} \text{ cm})^2}$$

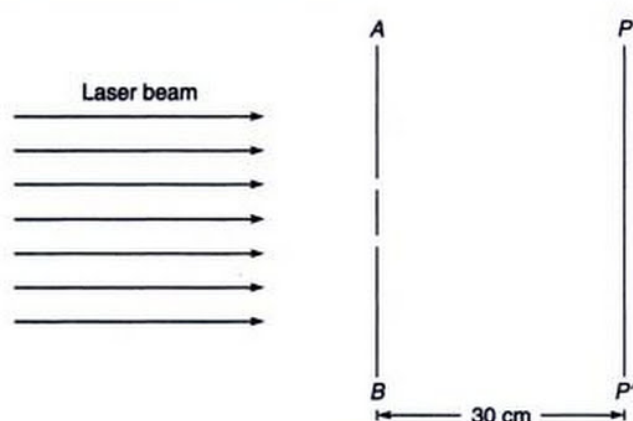


Fig. 1.13

$$\approx 3 \times 10^{-5} \text{ ergs} \approx 20 \text{ MeV}$$

Since the proton is bound inside the nucleus, the average of the potential energy, $\langle V \rangle$, must be negative and greater in magnitude than the kinetic energy. Therefore

$$-\langle V \rangle \geq 20 \text{ MeV}$$

which indeed gives the correct order of the potential energy. The uncertainty in momentum for the electron is again \hbar/r_0 ; however, since the rest mass of the electron is very much smaller than that of the proton, the velocity of the electron is very close to c and we have to use the extreme relativistic formula for the energy,

$$E = cp = \frac{c\hbar}{r_0} \\ \approx \frac{(3 \times 10^{10}) (1.05 \times 10^{-27})}{10^{-13} \times 1.6 \times 10^{-6}} \text{ MeV} \sim 200 \text{ MeV}$$

Although electrons do emerge from nuclei in β -decay, they seldom have energies exceeding a few million electron volts. Thus one does not expect the electron to be a basic constituent of the nucleus; the rare occasions when β -decay occurs may be attributed to the transformation of a neutron into a proton and an electron (and the neutrino) so that the electron is in fact created at the instant the decay occurs.

REFERENCES AND SUGGESTED READINGS

1. R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. III, Addison Wesley Publishing Co., Reading, Mass., 1965.
2. Isaac Newton, *Opticks*, Dover Publications, 1952. This is a reprint of the book first published in 1704.
3. W.B. Joyce and A. Joyce, 'Descartes, Newton and Snell's law', *Journal of Optical Society of America*, Vol. 66, (1), 1976.
4. A. W. Barton, *A Text Book on Light*, Longmans Green & Co., London, 1939.
5. C. Huygens, *Treatise on Light*, Dover Publications, 1962.
6. R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. II, Addison Wesley Publishing Co., Reading, Mass., 1965.
7. A. Einstein, *On a heuristic point of view concerning the production and transformation of light*, *Annalen der Physik*, 17, 132, 1905.
8. J. Stachel (Ed.), *Einstein's Miraculous Year: Five papers that changed the face of Physics*, Princeton University Press, Princeton, 1998; reprinted by Srishti Publishers, New Delhi.
9. G. Gamow, *Mr. Tompkins in Wonderland*, Cambridge University Press, Cambridge, 1940.
10. M. Born, *Atomic Physics*, Blackie & Son, London, 1962.
11. C.G. Shull, 'Neutron Diffraction: A General Tool in Physics' in *Current Problems in Neutron Scattering*, Proceedings of the Symposium held at CNEN Casaccia Center in September 1968, CNEN, Rome, 1970.
12. A. Rose, 'Quantum Effects in Human Vision', *Advances in Biological and Medical Physics*, Vol. V, Academic Press, 1957.
13. P.A.M. Dirac, *The Principles of Quantum Mechanics*, Oxford University Press, Oxford, 1958.
14. A. Tonomura, J. Endo, T. Matsuda, T. Kawasaki and H. Ezawa, 'Demonstration of single-electron build-up of an interference pattern', *American Journal of Physics*, 57 (2), 117, 1989.

PART 1

Geometrical Optics

This part (consisting of four chapters) is entirely based on geometrical optics and includes

- Ray tracing through graded-index media explaining in detail the phenomena of mirage and looming and also reflection from the ionosphere.
- Ray tracing through a system of lenses leading to various concepts used in the design of optical instruments.
- A detailed description of the matrix method in paraxial optics, which is extensively used in the industry.
- A study of aberrations of optical systems.

Chapter 2

Fermat's Principle and Its Applications

Now in the further development of science, we want more than just a formula. First we have an observation, then we have numbers that we measure, then we have a law which summarizes all the numbers. But the real glory of science is that we can find a way of thinking such that the law is evident. The first way of thinking that made the law about the behavior of light evident was discovered by Fermat in about 1650, and it is called the principle of least time, or Fermat's principle.

Feynman Lectures in Physics, Vol. I

Important Milestones

- 140 AD Greek physicist Claudius Ptolemy measured the angle of refraction in water for different angles of incidence in air and made a table of it.
- 1621 Although the above mentioned numerical table was made in 140 AD, it was only in 1621 that Willebrord Snell, a Dutch mathematician, discovered the law of refraction which is now known as Snell's law.
- 1637 Descartes derived the Snell's law; his derivation assumed corpuscular model of light.
- 1657 Pierre de Fermat enunciated his principle of 'least time' and derived Snell's law of refraction and showed that if the velocity of light in the second medium is less, the ray would bend towards the normal, contrary to what predicted by the 'corpuscular theory'.

2.1 INTRODUCTION

The study of the propagation of light in the realm of geometrical optics employs the concept of rays. To understand what a ray is, consider a circular aperture in front of a point source P as shown in Fig. 2.1. When the diameter of the aperture is quite large (~ 1 cm), then on the screen SS', one can see a patch of light with well-defined boundaries. When we start decreasing the size of the aperture, then at first the size of the patch starts decreasing, but when the size of the aperture becomes very small (≤ 0.1 mm) then the pattern obtained on SS' ceases to have well-defined boundaries. This phenomenon is known as diffraction and is a direct consequence of the finiteness of the wavelength (which is denoted by λ). In Chapters 16 and 17 we will discuss the phenomenon of diffraction in great detail and will show that the diffraction effects become smaller with the decrease in wavelength and indeed in the limit of $\lambda \rightarrow 0$, the diffraction effects will be absent, and even for extremely small diameters of the aperture, we will

obtain a well-defined shadow on the screen SS'; and therefore, in the zero wavelength limit one can obtain an infinitesimally thin pencil of light; this is called a ray. Thus, a ray defines the path of propagation of the energy in the limit of the wavelength going to zero. Since light has a wavelength of the order of 10^{-5} cm, which is small compared to the dimensions of normal optical instruments like lenses, mirrors, etc., one can, in many applications, neglect the finiteness of the wavelength. The field of optics under such an approximation (i.e., the neglect of the finiteness of the wavelength) is called geometrical optics.

The field of geometrical optics can be studied by using Fermat's principle which determines the path of the rays. According to this principle the ray will correspond to that path for which the time taken is an extremum in comparison to nearby paths, i.e., it is either a minimum or a maximum or stationary*. Let $n(x, y, z)$ represent the position dependent refractive index. Then

$$\frac{ds}{c/n} = \frac{n ds}{c}$$

* The entire field of classical optics (both geometrical and physical) can be understood from Maxwell's equations, and of course, Fermat's principle can be derived from Maxwell's equations (see Ref. 1, 2).

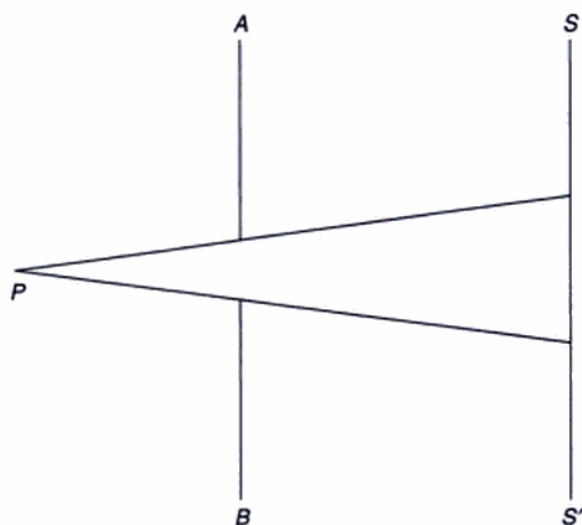


Fig. 2.1 The light emitted by the point source P is allowed to pass through a circular hole and if the diameter of the hole is very large compared to the wavelength of light then the light patch on the screen SS' has well defined boundaries.

will represent the time taken to traverse the geometric path ds in a medium of refractive index n . Here, c represents the speed of light in free space. Thus, if τ represents the total time taken by the ray to traverse the path AB along the curve C (see Fig. 2.2) then

$$\tau = \frac{1}{c} \sum_i n_i ds_i = \frac{1}{c} \int_{A \rightarrow B} n ds \quad (1)$$

where ds_i represents the i^{th} arc length and n_i the corresponding refractive index; the symbol $A \rightarrow B$ below the integral represents the fact that the integration is from the point A to B through the curve C. Let τ' be the time taken along the nearby path AC'B (shown as the dashed

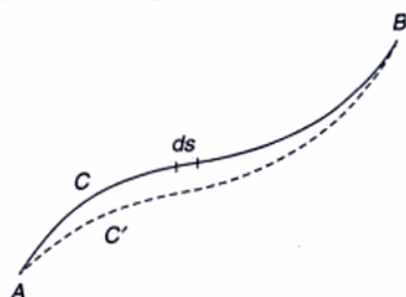


Fig. 2.2 If the path ACB represents the actual ray path then the time taken in traversing the path ACB will be an extremum in comparison to any nearby path AC'B.

curve in Fig. 2.2), and if ACB indeed represents the path of a ray, then τ will be either less than, greater than or equal to τ' for all nearby paths like AC'B. Thus according to Fermat's principle, out of the many paths connecting the two points, the light ray would follow that path for which the time taken is an extremum. Since c is a constant, one can alternatively define a ray as the path for which

$$\int_{A \rightarrow B} n ds \quad (2)$$

is an extremum^{**}; the above integral represents the optical path from A to B along C; i.e., the ray would follow the path for which

$$\delta \int_{A \rightarrow B} n ds = 0 \quad (3)$$

where the left-hand side represents the change in the value of the integral due to an infinitesimal variation of the ray path. We may mention here that according to the original statement of Fermat:

The actual path between two points taken by a beam of light is the one which is traversed in the least time.

The above statement is incomplete and slightly incorrect. The correct form is:

The actual ray path between two points is the one for which the optical path length is stationary with respect to variations of the path.

This is expressed by Eq. (3) and in this formulation, the ray paths may be maxima, minima or saddle points.

From the above principle one can immediately see that in a homogenous medium (i.e., in a medium whose refractive index is constant at each point), the rays will be straight lines because a straight line will correspond to a minimum value of the optical path connecting two points in the medium. Thus referring to Fig. 2.3, if A and B are two points in a homogenous medium, then the ray path will be along the straight line ACB because any nearby path like ADB or AEB will correspond to a longer time.

2.2 LAWS OF REFLECTION AND REFRACTION FROM FERMAT'S PRINCIPLE

We will now obtain the laws of reflection and refraction from Fermat's principle. Consider a plane mirror MN as

^{**} A nice discussion on the extremum principle has been given in Chapter 26 of Ref. 3.

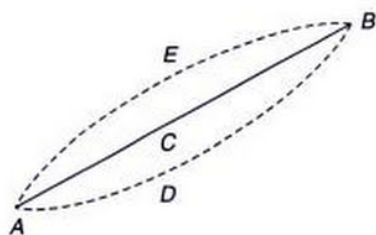


Fig. 2.3 Since the shortest distance between two points is along a straight line, light rays in a homogenous medium are straight lines; all nearby paths like AEB or ADB will take longer times.

shown in Fig. 2.4. To obtain the laws of reflection, we have to determine the path from A to B (via the mirror) which has the minimum optical path length. Since the path would lie completely in a homogenous medium, we need to minimize only the path length. Thus we have to find that path APB for which $AP + PB$ is a minimum. To find the position of P on the mirror, we drop a perpendicular from A on the mirror and let A' be a point on the perpendicular such that $AR = RA'$; thus $AP = PA'$ and $AQ = A'Q$ where AQB is another path adjacent to APB. Thus we have to minimize the length A'PB. Clearly, for A'PB to be a minimum, P must be on the straight line A'B. Thus the points A, A', P and B will be in the same plane and if we draw a normal PS at P then this normal will also lie in the same plane. Simple geometric considerations show that

$$\angle APS = \angle SPB$$

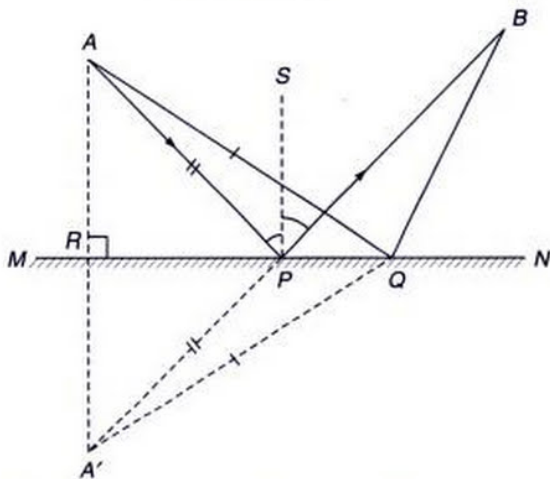


Fig. 2.4 The shortest path connecting the two points A and B via the mirror is along the path APB where the point P is such that AP, PS and PB are in the same plane and $\angle APS = \angle SPB$; PS being the normal to the plane of the mirror. The straight line path AB is also a ray.

Thus for minimum optical path length, the angle of incidence $i (= \angle APS)$ and the angle of reflection $r (= \angle SPB)$ must be equal and the incident ray, the reflected ray and the normal to the surface at the point of incidence on the mirror must be in the same plane. These form the laws of reflection. It should be pointed out that, in the presence of the mirror there will be two ray paths which will connect the points A and B; the two paths will be AB and APB. Fermat's principle tells us that whenever the optical path length is an extremum, we will have a ray, and thus, in general, there may be more than one ray path connecting two points.

To obtain the laws of refraction, let PQ be a surface separating two media of refractive indices n_1 and n_2 as shown in Fig. 2.5. Let a ray starting from the point A, intersect the interface at R and proceed to B along RB. Clearly, for minimum optical path length, the incident ray, the refracted ray and the normal to the interface must all lie in the same plane. To determine that point R for which the optical path length from A to B is a minimum, we drop perpendiculars AM and BN from A and B respectively on the interface PQ. Let $AM = h_1$, $BN = h_2$ and $MR = x$. Then since A and B are fixed, $RN = L - x$, where $MN = L$ is a fixed quantity. The optical path length from A to B, by definition, is

$$L_{op} = n_1 AR + n_2 RB \\ = n_1 \sqrt{x^2 + h_1^2} + n_2 \sqrt{(L - x)^2 + h_2^2} \quad (4)$$

To minimize this, we must have

$$\frac{dL_{op}}{dx} = 0$$

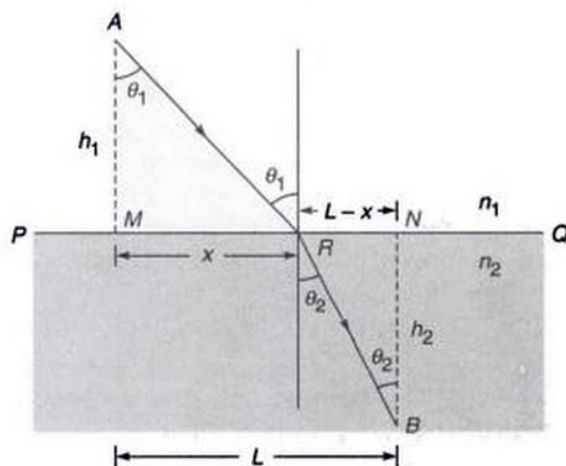


Fig. 2.5 A and B are two points in media of refractive indices n_1 and n_2 . The ray path connecting A and B will be such that $n_1 \sin \theta_1 = n_2 \sin \theta_2$.

$$\text{i.e., } \frac{n_1 x}{\sqrt{x^2 + h^2}} - \frac{n_2(L-x)}{\sqrt{(L-x)^2 + h^2}} = 0 \quad (5)$$

Further, as can be seen from Fig. 2.5

$$\sin \theta_1 = \frac{x}{\sqrt{x^2 + h_1^2}}$$

$$\text{and} \quad \sin \theta_2 = \frac{(L-x)}{\sqrt{(L-x)^2 + h_2^2}}$$

Thus Eq. (5) becomes

$$n_1 \sin \theta_1 = n_1 \sin \theta_2 \quad (6)$$

which is the Snell's law of refraction.

The laws of reflection and refraction form the basic laws for tracing light rays through simple optical systems, like a system of lenses and mirrors, etc.

Example 2.1 Consider a set of rays, parallel to the axis, and incident on a paraboloidal reflector. Show, by using Fermat's principle, that all the rays will pass through the focus of the paraboloid; a paraboloid is obtained by rotating a parabola about its axis. This is the reason why a paraboloidal reflector is used to focus parallel rays from a distant source, like in radio astronomy.

Solution: Consider a ray PQ, parallel to the axis of the parabola, incident at the point Q (see Fig. 2.6). In order to find the reflected ray, one has to draw a normal at the point Q and then draw the reflected ray. It can be shown from geometrical considerations that the reflected ray QS will always pass through the focus S. However, this procedure will be quite cumbersome and as we will show, the use of Fermat's principle leads us to the desired results immediately.

In order to use Fermat's principle we try to find out the ray connecting the focus S and an arbitrary point P (see Fig. 2.6). Let the ray path be PQ'S. According to Fermat's principle the ray path will correspond to a minimum value of PQ' + Q'S. From the point Q' we drop a perpendicular Q'L', on the directrix AB. From the definition of the parabola it follows that Q'L' = Q'S. Thus

$$PQ' + Q'S = PQ' + Q'L'$$

Let L be the foot of the perpendicular drawn from the point P on AB. Then, for PQ' + Q'L' to be a minimum, the point Q should lie on the straight line PQL, and thus the actual ray which connects the points P and S will be PQ + QS where PQ is parallel to the axis. Therefore, all rays parallel to the axis will pass through S and conversely, all rays emanating from the point S will become parallel to the axis after suffering a reflection.

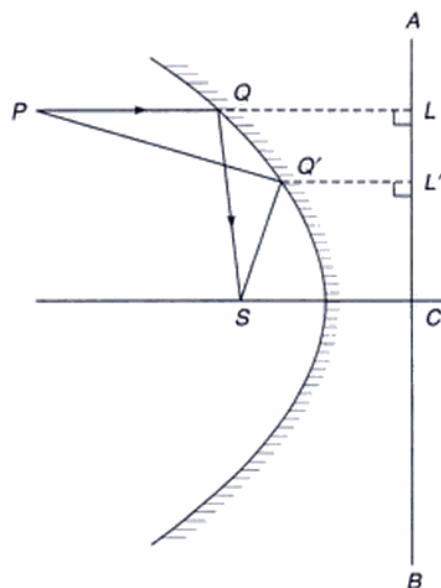


Fig. 2.6 All rays parallel to the axis of a paraboloidal reflector pass through the focus after reflection (the line ACB is the directrix). It is for this reason that antennas (for collecting electromagnetic waves) or solar collectors are often paraboloidal in shape.

Example 2.2 Consider an elliptical reflector whose foci are the points S_1 and S_2 (see Fig. 2.7). Show that all rays emanating from the point S_1 will pass through the point S_2 after undergoing a reflection.

Solution: Consider an arbitrary point P on the ellipse (see Fig. 2.7). It is well known that $S_1P + S_2P$ is a constant and therefore, all rays emanating from the point S_1 will pass through S_2 . Notice that here we have an example where the time taken by the ray is stationary, i.e., it is neither a maximum nor a minimum but has a constant value for all points lying on the mirror. As a corollary, we may note the following two points:

- (i) Excepting the rays along the axis, no other ray (emanating from either of the foci) will pass through an arbitrary point Q which lies on the axis.
- (ii) The above considerations will remain valid even for an ellipsoid of revolution obtained by rotating the ellipse about its major axis.

Because of the above mentioned property of elliptical reflectors, they are often used in laser systems. For example, in a ruby laser (see Chapter 23) one may have a scheme in which the laser rod and the flash lamp coincide with the focal lines of a cylindrical reflector of elliptical cross-section; such a configuration leads to an efficient transfer of energy from the lamp to the ruby rod.

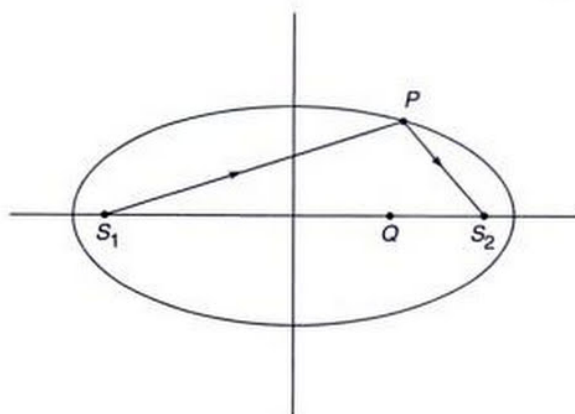


Fig. 2.7 All rays emanating from one of the foci of an ellipsoidal reflector will pass through the other focus.

Example 2.3 Consider a spherical refracting surface SPM separating two media of refractive indices n_1 and n_2 (see Fig. 2.8). The point C represents the center of the spherical surface SPM. Consider two points O and Q such that the points O, C and Q are in a straight line. Calculate the optical path length OSQ in terms of the distances x , y , r and the angle θ (see Fig. 2.8). Use Fermat's principle to find the ray connecting the two points O and Q. Also, assuming the angle θ to be small, determine the paraxial image of the point O.

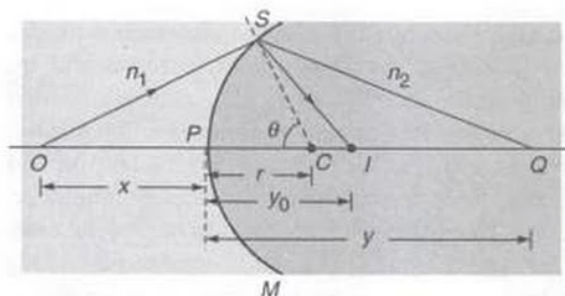


Fig. 2.8 SPM is a spherical refracting surface separating two media of refractive indices n_1 and n_2 . C represents the center of the spherical surface.

[Note: We reserve the symbol R to represent the radius of curvature of a spherical surface which will be positive (or negative) depending upon whether the center of curvature lies on the right (or left) of the point P. The quantity r represents the magnitude of the radius of curvature which, for Fig. 2.8, happens to be R . Similarly, the quantities x and y are the magnitudes of the distances; the sign convention is discussed later on in this problem.]

Solution: From the triangle SOC we have

$$OS = [(x+r)^2 + r^2 - 2(x+r)r \cos \theta]^{1/2}$$

$$= \left[x^2 + 2rx + 2r^2 - 2(xr + r^2) \left(1 - \frac{\theta^2}{2} \right) \right]^{1/2}$$

$$\approx x \left[1 + \frac{rx + r^2}{x^2} \theta^2 \right]^{1/2} \approx x + \frac{1}{2} r^2 \left(\frac{1}{r} + \frac{1}{x} \right) \theta^2$$

where we have assumed θ (measured in radians) to be small so that we may use the expression

$$\cos \theta = 1 - \frac{\theta^2}{2}$$

and also make a binomial expansion. Similarly, by considering the triangle SCQ we would have

$$SQ = y - \frac{1}{2} r^2 \left(\frac{1}{r} - \frac{1}{y} \right) \theta^2$$

Thus the optical path length OSQ is given by

$$L_{op} = n_1 OS + n_2 SQ \\ = (n_1 x + n_2 y) + \frac{1}{2} r^2 \left[\frac{n_1}{x} + \frac{n_2}{y} - \frac{n_2 - n_1}{r} \right] \theta^2 \quad (7)$$

For the optical path to be an extremum we must have

$$\frac{dL_{op}}{d\theta} = 0 = r^2 \left[\frac{n_1}{x} + \frac{n_2}{y} - \frac{n_2 - n_1}{r} \right] \theta \quad (8)$$

Thus, unless the quantity inside the square brackets is zero we must have $\theta = 0$ implying that the *only* ray connecting the points O and Q will be the straight line path OPQ which also follows from Snell's law because the ray OP hits the spherical surface normally and should proceed undeviated.

On the other hand, if the value of y was such that the quantity inside the square brackets was zero, i.e., if y was equal to y_0 such that

$$\frac{n_2}{y_0} + \frac{n_1}{x} = \frac{n_2 - n_1}{r} \quad (9)$$

then $dL_{op}/d\theta$ would vanish for *all* values of θ ; of course, θ is assumed to be small—which is the paraxial approximation. Now, if the point I corresponds to $PI = y_0$ (see Fig. 2.8) then *all* paths like OSI are allowed ray paths implying that *all* (paraxial) rays emanating from O will pass through I and I will therefore represent the paraxial image point. Obviously, all rays like OSI (which start from O and pass through I) take the *same* amount of time in reaching the point I.

We should mention that Eq. (9) is a particular form of the equation determining the paraxial image point

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \quad (10)$$

with the sign convention that all distances measured to the right of the point P are positive and those to its left negative. Thus $u = -x$, $v = +y$ and $r = +R$.

In order to determine whether the ray path OPQ corresponds to minimum time or maximum time or stationary, we must determine the sign of $d^2L_{op}/d\theta^2$ which is given by

$$\begin{aligned}\frac{d^2L_{op}}{d\theta^2} &= r^2 \left[\frac{n_1}{x} + \frac{n_2}{y} - \frac{n_2 - n_1}{r} \right] \\ &= r^2 n_2 \left[\frac{1}{y} - \frac{1}{y_0} \right]\end{aligned}$$

Obviously, if $y > y_0$ (i.e., the point Q is on the right of the paraxial image point I) $d^2L_{op}/d\theta^2$ is negative and the ray path OPQ corresponds to *maximum* time in comparison with nearby paths and conversely. On the other hand, if $y = y_0$, $d^2L_{op}/d\theta^2$ will vanish implying that the extremum corresponds to stationarity.

Alternatively, one can argue that if I is the paraxial image point of P then

$$n_1 OP + n_2 PI = n_1 OS + n_2 SI$$

Thus, when Q lies on the right of the point I, we have

$$\begin{aligned}n_1 OP + n_2 PQ &= n_1 OS + n_2 (SI - PI + PQ) \\ &= n_1 OS + n_2 (SI + IQ) \\ &> n_1 OS + n_2 SQ\end{aligned}$$

implying that the ray path OPQ corresponds to a maximum. Similarly, when Q lies on the left of the point I then the ray path OPQ corresponds to a minimum and when Q coincides with I, we have the stationarity condition.

Example 2.4 We again consider refraction at a spherical surface; however, the refracted ray is assumed to diverge away from the principal axis (see Fig. 2.9). Let us consider paraxial rays and let I be a point (on the axis) such that $n_1 OS - n_2 SI$ is independent of the point S. Thus, for paraxial rays, the quantity

$$n_1 OS - n_2 SI \quad \text{is independent of } \theta \quad (11)$$

and is an extremum. Let P be an arbitrary point in the second medium and we wish to find the ray path connecting the points O and P. For OSP to be an allowed ray path

$$L_{op} = n_1 OS + n_2 SP \text{ should be an extremum}$$

or

$$L_{op} = (n_1 OS - n_2 SI) + n_2 (SI + SP) \text{ should be an extremum}$$

where we have added and subtracted $n_2 SI$. Now, the point I is such that the first quantity is already an extremum thus, the quantity $SP + SI$ should be an extremum and therefore

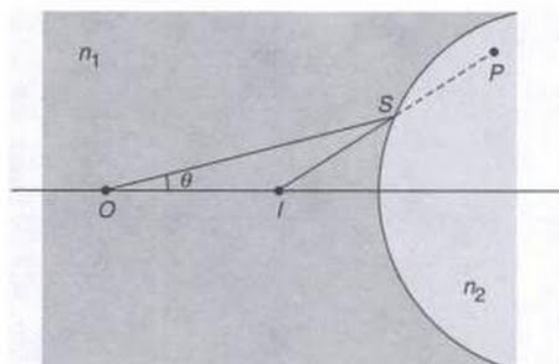


Fig. 2.9 The refracted ray is assumed to diverge away from the principal axis.

it should be a straight line. Thus the refracted ray must appear to come from the point I. We may therefore say that for a virtual image we must make the quantity

$$n_1 OS - n_2 SI \quad (12)$$

an extremum.

2.3 RAY PATHS IN AN INHOMOGENEOUS MEDIUM

In an inhomogeneous medium, the refractive index varies in a continuous manner and, in general, the ray paths are curved. For example, on a hot day, the air near the ground has a higher temperature than the air which is much above the surface. Since the density of air decreases with increase in temperature, the refractive index increases continuously as we go above the ground. This leads to the phenomenon known as *mirage*. We will use Snell's law (or Fermat's principle) to determine the ray paths in an inhomogeneous medium. We will restrict ourselves to the special case when the refractive index changes continuously along one direction only; we assume this direction to be along the x -axis.

The inhomogeneous medium can be thought of as a limiting case of a medium consisting of a continuous set of thin slices of media of different refractive indices—see Fig. 2.10(a). At each interface, the light ray satisfies Snell's law and one obtains [see Fig. 2.10(a)]

$$n_1 \sin \phi_1 = n_2 \sin \phi_2 = n_3 \sin \phi_3 = \dots \quad (13)$$

Thus, we may state that the product

$$n(x) \cos \theta(x) = n(x) \sin \phi(x) \quad (14)$$

is an invariant of the ray path; we will denote this invariant by $\tilde{\beta}$. The value of this invariant may be determined from the fact that if the ray initially makes an angle θ_1 (with the z -axis) at a point where the refractive index is n_1 , then the

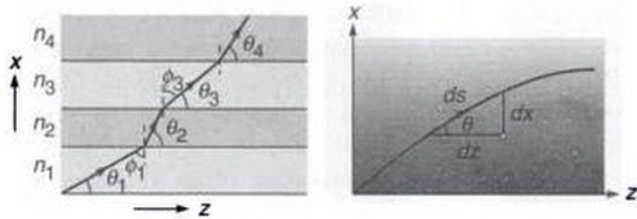


Fig. 2.10 (a) In a layered structure, the ray bends in such a way that the product $n_i \cos \theta_i$ remains constant. (b) For a medium with continuously varying refractive index, the ray path bends in such a way that the product $n(x) \cos \theta(x)$ remains constant.

value of $\tilde{\beta}$ is $n_1 \cos \theta_1$. Thus, in the limiting case of a continuous variation of refractive index, the piecewise straight lines shown in Fig. 2.10(a) form a continuous curve which is determined from the equation

$$n(x) \cos \theta(x) = n_1 \cos \theta_1 = \tilde{\beta} \quad (15)$$

implying that as the refractive index changes, the ray path bends in such a way that the product $n(x) \cos \theta(x)$ remains constant [see Fig. 2.10(b)]. Equation (15) can be used to derive the ray equation (see Sec. 2.4).

2.3.1 The Phenomenon of Mirage*

We are now in a position to qualitatively discuss the formation of a mirage. As mentioned earlier, on a hot day the refractive index continuously decreases as we go near the ground. Indeed, the refractive index variation can be approximately assumed to be of the form

$$n(x) \approx n_0 + kx \quad 0 < x < \text{few metres} \quad (16)$$

where n_0 is the refractive index of air at $x = 0$ (i.e., just above the ground) and k is a constant. The exact ray paths (see Example 2.8) are shown in Fig. 2.11.

We consider a ray which becomes horizontal at $x = 0$. At the eye position E ($x = x_e$), if the refractive index is n_e , and if at that point the ray makes an angle θ_e with the horizontal then

$$\tilde{\beta} = n_0 = n_e \cos \theta_e \quad (17)$$

Usually $\theta_e \ll 1$ so that

$$\frac{n_0}{n_e} = \cos \theta_e \approx 1 - \frac{1}{2} \theta_e^2$$

$$\Rightarrow \theta_e = \sqrt{2 \left(1 - \frac{n_0}{n_e} \right)} \quad (18)$$

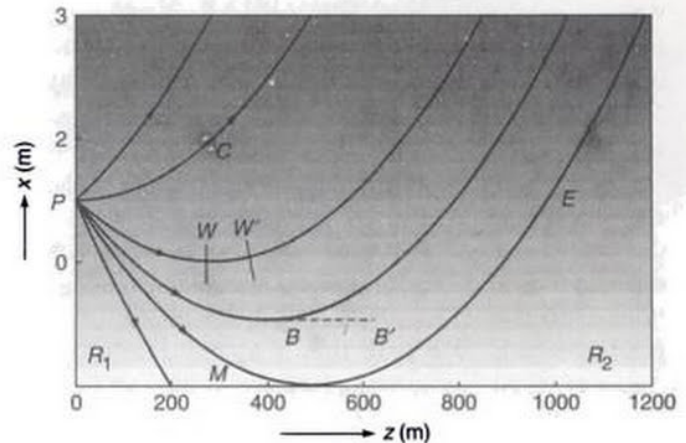


Fig. 2.11 Ray paths in a medium characterized by a linear variation of refractive index [see Eq. (16)] with $k \approx 1.234 \times 10^{-5} \text{ m}^{-1}$. The object point is at a height of 1.5 m and the curves correspond to $+0.2^\circ$, 0° , -0.2° , -0.28° , -0.3486° and -0.5° . The shading shows that the refractive index increases with x .

At constant air pressure

$$(n_0 - 1)T_0 \approx (n_e - 1)T_e \quad (19)$$

From Eq. (19) we get

$$1 - \frac{n_0 - 1}{n_e - 1} = 1 - \frac{T_e}{T_0}$$

or

$$\frac{n_e - n_0}{n_e} = \frac{n_e - 1}{n_e} \left(1 - \frac{T_e}{T_0} \right)$$

so that

$$\theta_e \approx \sqrt{2 \left(1 - \frac{1}{n_e} \right) \left(1 - \frac{T_e}{T_0} \right)} \quad (20)$$

On a typical hot day the temperature near road surface $T_0 \approx 323^\circ \text{K}$ ($= 50^\circ \text{C}$) and, about 1.5 m above the ground, $T_e \approx 303^\circ \text{K}$ ($= 30^\circ \text{C}$). Now, at 30°C , $n_e = 1.00026$ giving $\theta_e \approx 5.67 \times 10^{-3}$ radians $\approx 0.325^\circ$. In Fig. 2.11 we have shown rays emanating (at different angles) from a point P which is 1.5 m above the ground; thus each ray has a specified value of the invariant $\tilde{\beta}$ ($= n_1 \cos \theta_1$). The figure shows that when the object point P and the observation point E are close to the ground, the only ray path connecting points P and E will be along the curve PME and that a ray emanating horizontally from the point P will propagate in the upward direction as PC as shown in figure. Thus, in such a condition, the eye at E will see the mirage and *not* see the object directly at P. We also find that there is a region R_2

*For more details, see Refs 4–8.

where none of the rays (emanating from the point P) reaches; thus, an eye in this region can neither see the object nor its image. This is therefore called the *shadow region*. Furthermore, there is also a region R_1 where only the object is directly visible and the virtual image is not seen.

We should mention here that the *bending up* of the ray after it becomes parallel to the z -axis cannot be directly inferred from Eq. (15) because at such a point, $\theta = 0$ and one may expect the ray to proceed horizontally beyond the turning point as shown by a dotted line in Fig. 2.11; the point at which $\theta = 0$ is known as the *turning point*. However, from considerations of symmetry and from the reversibility of ray paths, it immediately follows that the ray path should be symmetrical about the turning point and hence bend up. Physically, the bending of the ray can be understood by considering a small portion of a wave front

such as W (see Fig. 2.11); the upper edge will travel with a smaller speed in comparison with the lower edge, and this will cause the wave front to tilt (see W') making the ray to bend. Furthermore, a straight line path like BB' does not correspond to an extremum value of the optical path.

We next consider a refractive index variation which saturates to a constant value as $x \rightarrow \infty$:

$$n^2(x) = n_0^2 + n_2^2 (1 - e^{-\alpha x}); \quad x > 0 \quad (21)$$

where n_0 , n_2 and α are constants and once again, x represents the height above the ground. The refractive index at $x = 0$ is n_0 and for large values of x , it approaches $(n_0^2 + n_2^2)^{1/2}$. The exact ray paths are obtained by solving the ray equation (see Example 2.10) and are shown in Figs 2.12 and 2.13; they correspond to the following values of various parameters:

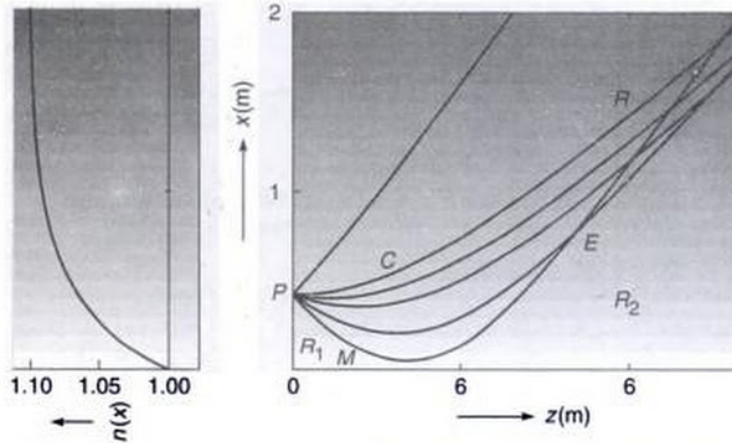


Fig. 2.12 Ray paths in a medium characterised by Eqs. (21) and (22). The object point is at a height of $1/\alpha (\approx 0.43\text{m})$ and the curves correspond to θ_1 (the initial launch angle) = $+\pi/10$, 0 , $-\pi/60$, $-\pi/30$, $-\pi/15$ and $-\pi/10$. The shading shows that the refractive index increases with x .

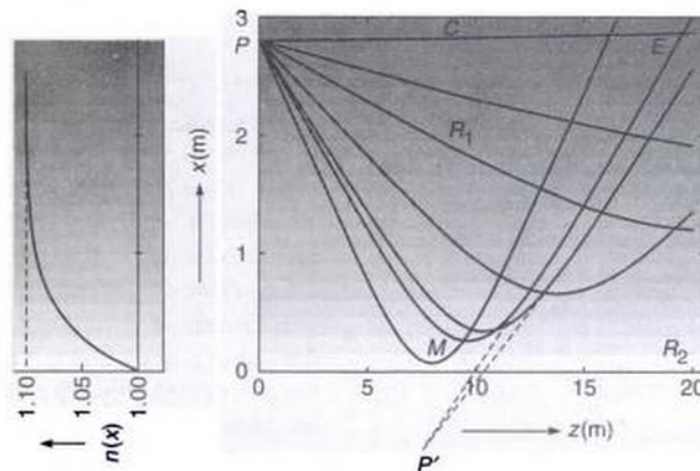


Fig. 2.13 Ray paths in a medium characterized by Eqs. (21) and (22). The object point is at a height of 2.8 m and the curves correspond to θ_1 (the initial launch angle) = 0 , $-\pi/60$, $-\pi/30$, $-\pi/16$, $-\pi/11$, $-\pi/10$ and $-\pi/8$. The shading shows that the refractive index increases with x .

$$\text{and} \quad \begin{aligned} n_0 &= 1.000233, \quad n_2 = 0.45836 \\ \alpha &= 2.303 \text{ m}^{-1} \end{aligned} \quad (22)$$

The actual values of the refractive index for parameters given by the above equation is not very realistic—nevertheless, it allows us to understand qualitatively the ray paths in a graded index medium. Figures 2.12 and 2.13 show the ray paths emanating from points that are 0.43 m and 2.8 m above the ground respectively. In Fig. 2.12, the point P corresponds to a value of the refractive index equal to 1.06455 ($= n_1$) and different rays correspond to different values of θ_1 , the angle that the ray makes with the z-axis at the point P. From Fig. 2.12 we again see that when the object point P and the observation point E are close to the ground, the only ray path connecting points P and E will be along the curve PME and that a ray emanating horizontally from the point P will propagate in the upward direction, shown as PC in Fig. 2.12. Thus, in such a condition, the eye at E will see the mirage and *not* see the object directly at P. However, if points P and E are much above the ground (see Fig. 2.13), the eye will see the object almost directly (because of rays like PCE) and will also receive rays appearing to emanate from points like P'. It may be readily seen that different rays do not appear to come from the same point and hence the reflected image seen will have considerable aberrations. Once again, there is a *shadow region* R_2 where none of the rays (emanating from the point P) reaches there; thus, an eye in this region can neither see the object nor its image.

Example 2.5 As an example, for an object shown in Fig. 2.12, let us calculate the angle at which the ray should be launched so that it becomes horizontal at $x = 0.2$ m. Now,

$$\text{at } x = 0.2 \text{ m}, \quad n(x) = 1.03827$$

Thus, if θ_1 represents the angle that the ray makes with the z-axis at the point P (see Fig. 2.12) then

$$n_1 \cos \theta_1 = 1.03827 \times \cos 0$$

implying

$$\theta_1 \approx 13^\circ$$

Further, for the ray which becomes horizontal at $x = 0.2$ m the value of the invariant is given by

$$\tilde{\beta} \approx 1.03827$$

Example 2.6 In Fig. 2.13, the object point corresponds to $x = 2.8$ m where $n(x) \approx 1.1$. Thus for a ray launched with $\theta_1 = -\pi/8$,

$$\tilde{\beta} = 1.1 \cos \theta_1 = 1.01627$$

Thus if the ray becomes horizontal at $x = x_2$ then

$$n(x_2) = \tilde{\beta} = 1.01627$$

and

$$\begin{aligned} x_2 &= -\frac{1}{\alpha} \ln \left[1 - \frac{n^2(x_2) - n_0^2}{n_2^2} \right] \\ &\approx 0.073 \text{ m} \end{aligned}$$

2.3.2 The Phenomenon of Looming

The formation of mirage discussed above occurs due to increase in the refractive index of air above the hot surface. On the other hand, above cold sea water, the air near the water surface is colder than the air above it and hence there is an opposite temperature gradient. A suitable refractive index variation for such a case can be written as:

$$n^2(x) = n_0^2 + n_2^2 e^{-\alpha x} \quad (23)$$

The equation describing the ray path is discussed in Problem 2.13. We assume the values of n_0 , n_2 and α to be given by Eq. (22). For an object point P at a height of 0.5 m, the ray paths are shown in Fig. 2.14. If the eye is at E, then it will receive rays appearing to emanate from P'. Such a

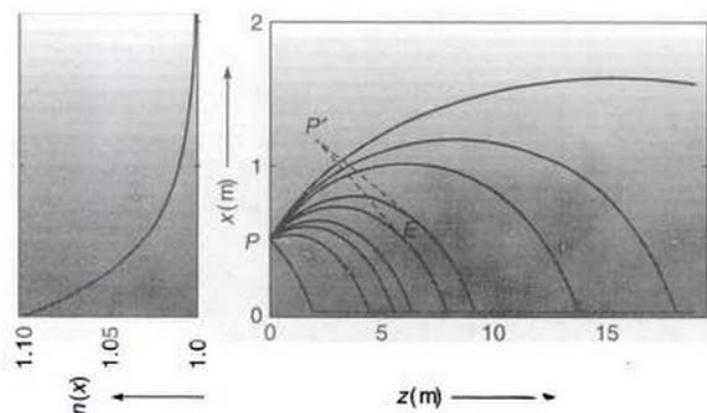


Fig. 2.14 Ray paths corresponding to the refractive index distribution given by Eq. (23) for an object at a height of 0.5 m; the values of n_0 , n_2 and α are given by Eq. (22).

phenomenon in which the object appears to be above its actual position is known as *looming*; it is commonly observed in viewing ships over cold sea waters. Moreover, since no other rays emanating from P reach A, the object cannot be observed directly.

2.3.3 The Graded Index Atmosphere

One of the interesting phenomena associated with imaging in a graded index medium is the noncircular shape of the setting or the rising sun (see Fig. 2.15). This can easily be understood in the following manner. The refractive index of the air gradually decreases as we move outwards. If we approximate the continuous refractive index gradient by a finite number of layers (each layer having a specific refractive index) then the ray will bend in a way similar to that shown in Fig. 2.16. Thus the sun (which is actually at S) appears to be in the direction of S'. It is for this reason that the setting sun appears flattened and also leads to the fact that the days are usually about 5 minutes longer than they would have been in the absence of the atmosphere.



Fig. 2.15 The non-circular shape of the setting sun.

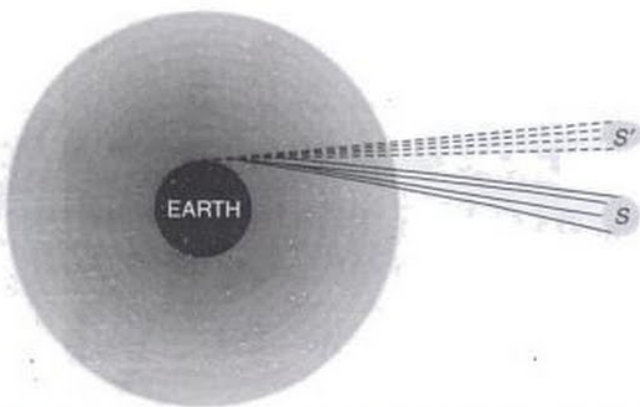


Fig. 2.16 Because of refraction, light from S appears to come from S'.

Obviously, if we were on the surface of the moon, the rising or the setting sun would not only look white but also circular in shape!

2.4 THE RAY EQUATION AND ITS SOLUTIONS

In this section, we will derive the ray equation, the solution of which will give the precise ray paths in an inhomogeneous medium. We will restrict ourselves to the special case when the refractive index changes continuously along only one direction, which we assume to be along the x -axis. This medium can be thought of as the limiting case of a medium comprising of a continuous set of thin slices of media of different refractive indices. As discussed earlier, for a continuously varying refractive index, the product $n(x) \cos \theta(x)$ is an invariant of the ray path which we denote by $\tilde{\beta}$:

$$n(x) \cos \theta(x) = \tilde{\beta} \quad (24)$$

Furthermore, for a continuous variation of refractive index, the piecewise straight lines shown in Fig. 2.10(a) forms a continuous curve as in Fig. 2.10(b). If ds represents the infinitesimal arc length along the curve, then

$$(ds)^2 = (dx)^2 + (dz)^2$$

or

$$\left(\frac{ds}{dz}\right)^2 = \left(\frac{dx}{dz}\right)^2 + 1 \quad (25)$$

Now, if we refer to Fig. 2.10(b), we find that

$$\frac{dz}{ds} = \cos \theta = \frac{\tilde{\beta}}{n(x)} \quad (26)$$

Thus Eq. (25) becomes

$$\left(\frac{dx}{dz}\right)^2 = \frac{n^2(x)}{\tilde{\beta}^2} - 1 \quad (27)$$

For a given $n(x)$ variation, Eq. (27) can be integrated to give the ray path $x(z)$; however, it is often more convenient to put Eq. (27) in a slightly different form by differentiating it with respect to z :

$$2 \frac{dx}{dz} \frac{d^2x}{dz^2} = \frac{1}{\tilde{\beta}^2} \frac{dn^2}{dx} \frac{dx}{dz}$$

or

$$\frac{d^2x}{dz^2} = \frac{1}{2\tilde{\beta}^2} \frac{dn^2}{dx} \quad (28)$$

which represents the rigorously correct ray equation when the refractive index depends only on the x -coordinate.

Example 2.7 As a simple application of Eq. (28), let us consider a homogeneous medium for which $n(x)$ is a constant. In such a case, the RHS of Eq. (28) is zero and one obtains

$$\frac{d^2x}{dz^2} = 0$$

Integrating the above equation twice with respect to z , we obtain

$$x = Az + B$$

which is the equation of a straight line, as it ought to be in a homogeneous medium.

Example 2.8 We next consider the ray paths in a medium characterized by the following refractive index variation

$$n(x) = n_0 + kx \quad (29)$$

For the above profile, the ray equation [Eq. (28)] takes the form

$$\frac{d^2x}{dz^2} = \frac{1}{2\tilde{\beta}^2} \frac{dn^2}{dx} = \frac{k}{\tilde{\beta}^2} [n_0 + kx]$$

or

$$\frac{d^2X}{dz^2} = \kappa^2 X(z) \quad (30)$$

where

$$X \equiv x + \frac{n_0}{k} \text{ and } \kappa = \frac{k}{\tilde{\beta}} \quad (31)$$

Thus the ray path is given by

$$x(z) = -\frac{n_0}{k} + C_1 e^{\kappa z} + C_2 e^{-\kappa z} \quad (32)$$

where the constants C_1 and C_2 are to be determined from initial conditions. We assume that at $z = 0$, the ray is launched at $x = x_1$ making an angle θ_1 with the z -axis; thus

$$x(z = 0) = x_1$$

and

$$\left. \frac{dx}{dz} \right|_{z=0} = \tan \theta_1$$

Elementary manipulations would give us

$$C_1 = \frac{1}{2} \left[x_1 + \frac{1}{k} (n_0 + n_1 \sin \theta_1) \right] \quad (33)$$

and

$$C_2 = \frac{1}{2} \left[x_1 + \frac{1}{k} (n_0 - n_1 \sin \theta_1) \right] \quad (34)$$

where $n_1 = n_0 + kx_1$ represents the refractive index at $x = x_1$ and we have used the fact that

$$\tilde{\beta} = n_1 \cos \theta_1 \quad (35)$$

Figure 2.11 shows the ray paths as given by Eq. (32) with $x_1 = 1.5$ m, $n_1 = 1.00026$ and $k = 1.234 \times 10^{-5} \text{ m}^{-1}$.

2.4.1 Ray Paths in Parabolic Index Media

We consider a parabolic index medium characterized by the following refractive index distribution:

$$n^2(x) = n_1^2 - \gamma^2 x^2 \quad (36)$$

We will use Eq. (27) to determine the ray paths. Equation (27) can be written as

$$\int \frac{dx}{\sqrt{n^2(x) - \tilde{\beta}^2}} = \pm \frac{1}{\tilde{\beta}} \int dz \quad (37)$$

Substituting for $n^2(x)$, we get

$$\int \frac{dx}{\sqrt{x_0^2 - x^2}} = \pm \Gamma \int dz \quad (38)$$

where

$$x_0 = \frac{1}{\gamma} \sqrt{n_1^2 - \tilde{\beta}^2} \quad (39)$$

and

$$\Gamma = \frac{\gamma}{\tilde{\beta}} \quad (40)$$

Writing $x = x_0 \sin \theta$ and carrying out the straightforward integration we get

$$x = \pm x_0 \sin [\Gamma(z - z_0)] \quad (41)$$

We can always choose the origin such that $z_0 = 0$ so that the general ray path would be given by

$$x = \pm x_0 \sin \Gamma z \quad (42)$$

We may mention here that we could have also used Eq. (28) to obtain the ray path. Now, in an optical waveguide the refractive index distribution is usually written in the form*:

$$\left. \begin{aligned} n^2(x) &= n_1^2 \left[1 - 2\Delta \left(\frac{x}{a} \right)^2 \right], & |x| < a \text{ core} \\ &= n_2^2 = n_1^2 (1 - 2\Delta), & |x| < a \text{ cladding} \end{aligned} \right\} \quad (43)$$

*Ray paths in such media are of tremendous importance as they readily lead to very important results for parabolic index fibers which are extensively used in fiber-optic communication systems (see Sec. 24.7).

The region $|x| < a$ is known as the core of the waveguide and the region $|x| > a$ is usually referred to as the cladding. Thus

$$\gamma = \frac{n_1 \sqrt{2\Delta}}{a} \quad (44)$$

In a typical parabolic index fiber,

$$n_1 = 1.5, \quad \Delta = 0.01, \quad a = 20 \mu\text{m} \quad (45)$$

giving

$$n_2 \approx 1.485$$

and

$$\gamma = 1.0607 \times 10^4 \text{ m}^{-1}$$

Typical ray paths for different values of θ_1 are shown in Fig. 2.17. Obviously, the rays will be guided in the core if $n_2 < \tilde{\beta} < n_1$. When $\tilde{\beta} = n_2$, the ray path will become horizontal at the core-cladding interface. For $\tilde{\beta} < n_2$, the ray will be incident at the core-cladding interface at an angle and the ray will be refracted away. Thus, we may write

$$\begin{aligned} n_2 < \tilde{\beta} < n_1 &\Rightarrow \text{Guided rays} \\ \tilde{\beta} < n_2 &\Rightarrow \text{Refracting rays} \end{aligned} \quad (46)$$

In Fig. 2.17, the ray paths shown correspond to

$$z_0 = 0 \quad \text{and} \quad \theta_1 = 4^\circ, 8.13^\circ \text{ and } 20^\circ;$$

the corresponding values of $\tilde{\beta}$ are approximately 1.496 ($> n_2$), 1.485 ($= n_2$) and 1.410 ($< n_2$)—the last ray undergoes refraction at the core-cladding interface. It may be readily seen that the periodical length z_p of the sinusoidal path is given by

$$z_p = \frac{2\pi}{\Gamma} = \frac{2\pi a \cos \theta_1}{\sqrt{2\Delta}} \quad (47)$$

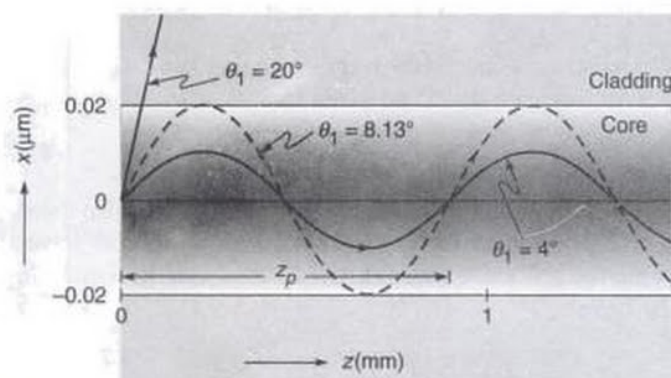


Fig. 2.17 Typical ray paths in a parabolic index medium for parameters given by Eq.(45) for $\theta_1 = 4^\circ, 8.13^\circ$ and 20° .

Thus for the two rays shown in Fig. 2.17 (with $\theta_1 = 4^\circ$ and 8.13°) the values of z_p would be 0.8864 mm and 0.8796 mm respectively. Indeed, in the paraxial approximation, $\cos \theta_1 \approx 1$ and all rays have the same periodic length. In Fig. 2.18, we have plotted typical paraxial ray paths for rays launched along the z -axis. Different rays (shown in the figure) correspond to different values of $\tilde{\beta}$.

Four interesting features may be noted:

- (i) In the paraxial approximation ($\tilde{\beta} = n_1$) all rays launched horizontally come to a focus at a particular point. Thus the medium acts as a converging lens of focal length given by:

$$f = \frac{\pi}{2} \frac{a}{\sqrt{2\Delta}} \quad (48)$$
- (ii) Rays launched at different angles with the axis (see, for instance, the rays emerging from point P) get trapped in the medium and hence the medium acts like a 'guide'. Indeed such media are referred to as optical waveguides and their study forms a subject of great contemporary interest.
- (iii) Ray paths would be allowed only in the region where $\tilde{\beta}$ is less than or equal to $n(x)$ [see Eq. (26)]. Further, dx/dz would be zero (i.e., the ray would become parallel to the z -axis) when $n(x)$ equals $\tilde{\beta}$; this immediately follows from Eq. (27).
- (iv) The rays periodically focus and defocus as shown in Fig. 2.18. In the paraxial approximation, all rays emanating from P will focus at Q and if we refer to our discussion in Example 2.3, all rays must take the same time to go from P to Q. Physically, although the ray PLQ traverses a larger path in comparison to PMQ, it does so in a medium of 'lower' average refractive index—thus the greater path length is compensated for by a greater 'average speed' and hence all rays take the same time to propagate through a certain distance of the waveguide (see Sec. 2.4.2 for exact calculation). It is for this reason that parabolic index waveguides are extensively used in fiber-optic communication systems (see Sec. 24.7).

We may mention here that Gradient-Index (GRIN) lenses, characterized by parabolic variation of refractive index in the transverse direction, are now commercially available and find many applications. For example a GRIN lens can be used to couple the output of a laser diode to an optical fiber; the length of such a GRIN lens would be $z_p/4$ (see Fig. 2.18); typically $z_p \approx$ few cm and the diameter of the lens would be few millimeters. Such small size lenses

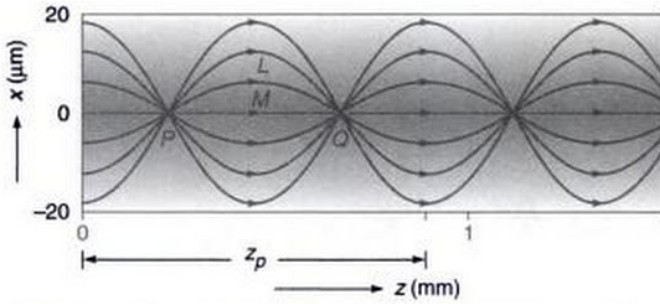


Fig. 2.18 Paraxial ray paths in a parabolic index medium. Notice the periodic focussing and defocussing of the beam.

find many applications. Similarly, a GRIN lens of length $z_p/2$ can be used to transfer collimated light from one end of the lens to the other.

2.4.2 Transit Time Calculations in a Parabolic Index Waveguide

In this section we will calculate the time taken by a ray to traverse a certain length through a parabolic index waveguide as described by Eq. (36). Such a calculation is of considerable importance in fiber optic communication systems (see Sec. 24.7). As shown in Sec. 2.4.1, the ray path (inside the core) is given by

$$x = x_0 \sin \Gamma z \quad (49)$$

where x_0 and Γ have been defined through Eqs. (39) and (40). Let $d\tau$ represent the time taken by a ray to traverse the arc length ds [see Fig. 2.10 (b)]:

$$d\tau = \frac{ds}{c/n(x)} \quad (50)$$

where c is the speed of light in free space. Since

$$n(x) \frac{dz}{ds} = \tilde{\beta}$$

[see Eq. (26)] we may write Eq. (50) as

$$\begin{aligned} d\tau &= \frac{1}{c\tilde{\beta}} n^2(x) dz \\ &= \frac{1}{c\tilde{\beta}} [n_1^2 - \gamma^2 x^2] dz \end{aligned}$$

$$\text{or} \quad d\tau = \frac{1}{c\tilde{\beta}} [n_1^2 - \gamma^2 x_0^2 \sin^2 \Gamma z] dz \quad (51)$$

where in the last step we have used Eq. (49). Thus if $\tau(z)$ represents the time taken by the ray to traverse a distance z along the waveguide then

$$\begin{aligned} \tau(z) &= \frac{n_1^2}{c\tilde{\beta}} \int_0^z dz - \frac{\gamma^2 x_0^2}{c\tilde{\beta}} \int_0^z \frac{1 - \cos(2\Gamma z)}{2} dz \\ &= \frac{1}{c\tilde{\beta}} \left[n_1^2 - \frac{1}{2} \gamma^2 x_0^2 \right] z + \frac{\gamma^2 x_0^2}{2c\tilde{\beta}} \frac{1}{2\Gamma} \sin 2\Gamma z \end{aligned}$$

$$\text{or} \quad \tau(z) = \frac{1}{2c\tilde{\beta}} [n_1^2 + \tilde{\beta}^2] z + \frac{(n_1^2 - \tilde{\beta}^2)}{4c\gamma} \sin 2\Gamma z \quad (52)$$

where we have used Eq. (39). When $\tilde{\beta} = n_1$ (which corresponds to the ray along the z -axis)

$$\tau(z) = \frac{z}{c/n_1} \quad (53)$$

which is what we should have expected as the ray will *always* travel with speed c/n_1 . For large values of z , the second term on the RHS of Eq. (52) would make a negligible contribution to $\tau(z)$ and we may write

$$\tau(z) \approx \frac{1}{2c} \left[\tilde{\beta} + \frac{n_1^2}{\tilde{\beta}} \right] z \quad (54)$$

Now, if a pulse of light is incident on one end of the waveguide, it would in general excite all rays and since different rays take different amounts of time, the pulse will get temporally broadened. Thus, for a parabolic index waveguide, this broadening will be given by

$$\Delta\tau = \tau(\tilde{\beta} = n_2) - \tau(\tilde{\beta} = n_1)$$

$$\text{or} \quad \Delta\tau = \frac{z}{2c} \frac{(n_1 - n_2)^2}{n_2} = \frac{zn_2}{2c} \Delta^2 \quad (55)$$

where in the last step we have assumed

$$\Delta \equiv \frac{n_1^2 - n_2^2}{2n_1^2} \approx \frac{n_1 - n_2}{n_2} \quad (56)$$

For the fiber parameters given by Eq. (45), we get

$$\Delta\tau = 0.25 \text{ ns/km} \quad (57)$$

We will use this result in Chapter 24.

Example 2.9 We next consider the ray paths in a medium characterized by the following refractive index variation

$$\begin{aligned} n^2(x) &= n_1^2 \quad x < 0 \\ &= n_1^2 - gx \quad x > 0 \end{aligned} \quad (58)$$

Thus, in the region $x > 0$, $n^2(x)$ decreases linearly with x and Eq. (28) takes the form

$$\frac{d^2x}{dz^2} = -\frac{g}{2\tilde{\beta}^2}$$

The general solution of which is given by

$$x(z) = -\frac{g}{4\tilde{\beta}^2} z^2 + K_1 z + K_2 \quad (59)$$

Consider a ray incident on the origin ($x = 0, z = 0$) as shown in Fig. 2.19. Thus

$$K_2 = 0 \quad \text{and} \quad \tilde{\beta} = n_1 \cos \theta_1 \quad (60)$$

Further,

$$\left. \frac{dx}{dz} \right|_{z=0} = K_1 = \tan \theta_1 \quad (61)$$

Thus the ray path will be given by

$$\left. \begin{aligned} &= (\tan \theta_1) z & z < 0 \\ x(z) &= -\frac{gz}{4\tilde{\beta}^2} (z - z_0) & 0 < z < z_0 \\ &= -\frac{gz_0}{4\tilde{\beta}^2} (z - z_0) & z > z_0 \end{aligned} \right\} \quad (62)$$

where

$$z_0 = \frac{2n_1^2}{g} \sin 2\theta_1$$

Thus in the region $0 < z < z_0$, the ray path is a parabola. Typical ray paths are shown in Fig. 2.19, the calculations corresponds to

$$n_1 = 1.5, \quad g = 0.1 \text{ m}^{-1}$$

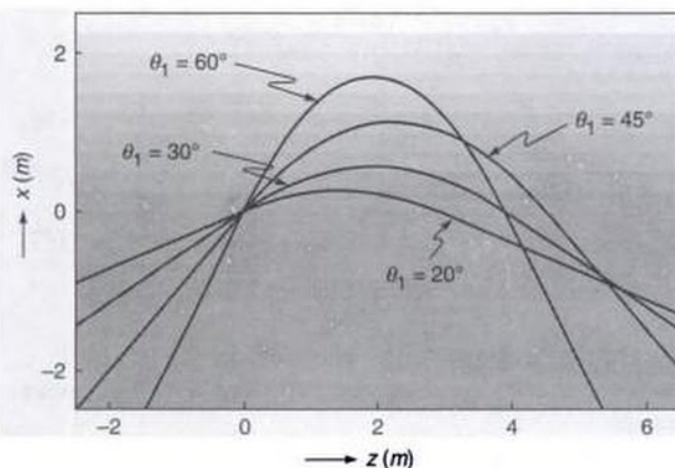


Fig. 2.19 Parabolic ray paths (corresponding to $\theta_1 = 20^\circ, 30^\circ, 45^\circ$ and 60°) in a medium characterized by refractive index variation given by Eq.(58). The ray paths in the region $x < 0$ are straight lines.

and different rays corresponds to

$$\theta_1 = \frac{\pi}{9}, \frac{\pi}{6}, \frac{\pi}{4} \quad \text{and} \quad \frac{\pi}{3}$$

2.4.3 Reflections from the Ionosphere

The ultraviolet rays in the solar radiation results in the ionization of the constituent gases in the atmosphere resulting in the formation of what is known as the ionosphere. The ionization is almost negligible below a height of about 60 km. Because of the presence of the free electrons (in the ionosphere), the refractive index is given by (see Eq. (76) of Chapter 6):

$$n^2(x) = 1 - \frac{N_e(x) q^2}{m\epsilon_0 \omega^2} \quad (63)$$

where

$N_e(x)$ represents the number of electrons/unit volume in m^{-3}

x represents the height above the ground in meters

ω represents the angular frequency of the electromagnetic wave

$q \approx 1.60 \times 10^{-19} \text{ C}$ represents the charge of the electron

$m = 9.11 \times 10^{-31} \text{ kg}$ represents the mass of the electron, and

$\epsilon_0 \approx 8.854 \times 10^{-12} \times 10^{-12} \text{ C}^2/\text{N-m}^2$ represents the dielectric permittivity of vacuum

Thus as the electron density starts increasing from 0 (beyond the height of 60 km) the refractive index starts decreasing and the ray paths would be similar to that described in Example 2.9.

We should mention that if n_T represents the refractive index at the turning point (where the ray becomes horizontal) then (see Fig. 2.20)

$$\tilde{\beta} = \cos \theta_1 = n_T \quad (64)$$

Thus if an electromagnetic signal is sent from the point A (at an angle θ_1) is received at the point B, one can determine the refractive index (and hence the electron density) of the ionospheric layer where the beam has undergone the reflection. This is how the short wave radio broadcasts ($\lambda \approx 20 \text{ m}$) sent at a particular angle from a particular city (say London) would reach another city (say New Delhi) after undergoing reflection from the ionosphere. Further, for normal incidence, $\theta = \pi/2$ and $n_T = 0$ implying

$$N_e(x_T) = \frac{m\epsilon_0 \omega^2}{q^2} \quad (65)$$

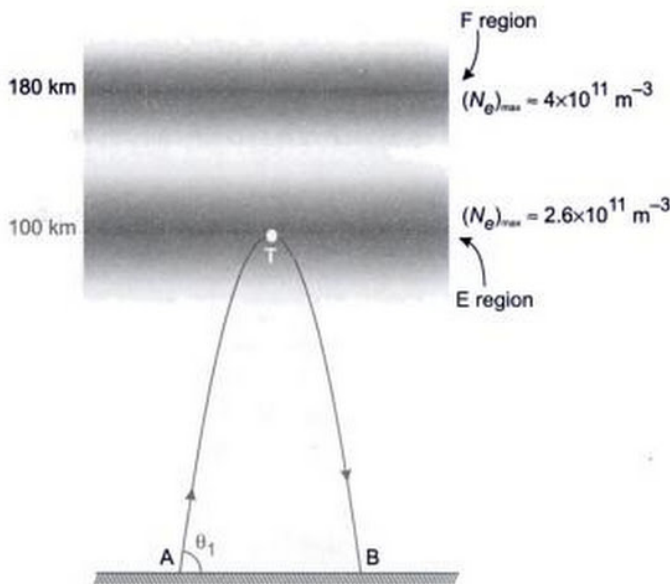


Fig. 2.20 Reflection from the *E* region of the ionosphere. The point *T* represents the turning point. The shading shows the variation of electron density.

In a typical experiment, an electromagnetic pulse (of frequency between 0.5 to 20 MHz) is sent vertically upwards and if the echo is received after a delay of Δt seconds, then

$$\Delta t \approx \frac{2h}{c} \quad (66)$$

where h represents the height at which it undergoes reflection. Thus if electromagnetic pulse is reflected from the *E* layer of ionosphere (which is at a height of about 100 km), the echo will be received after about 670 μ s. Alternatively, by measuring the delay Δt , one can determine the height (at which the pulse gets reflected) from the following relation

$$h = \frac{c}{2} \Delta t \quad (67)$$

In Fig. 2.21 we have plotted the frequency dependence of the equivalent height of reflection (as obtained from the delay time of echo) from the *E* and *F* regions of the ionosphere. From the figure we find that at $\nu = 4.6 \times 10^6$ Hz, echoes suddenly disappear from the 100 km height. Thus,

$$\begin{aligned} N_e(100 \text{ km}) &\approx \frac{m\epsilon_0 (2\pi\nu)^2}{q^2} \\ &\approx \frac{9.11 \times 10^{-31} \times 8.854 \times 10^{-12} \times (2\pi \times 4.6 \times 10^6)^2}{(1.6 \times 10^{-19})^2} \\ &\approx 2.6 \times 10^{11} \text{ electrons/m}^3 \end{aligned}$$

If we further increase the frequency, the echoes appear from the *F* region of the ionosphere. For more details of the studies on the ionosphere, the reader is referred to one of

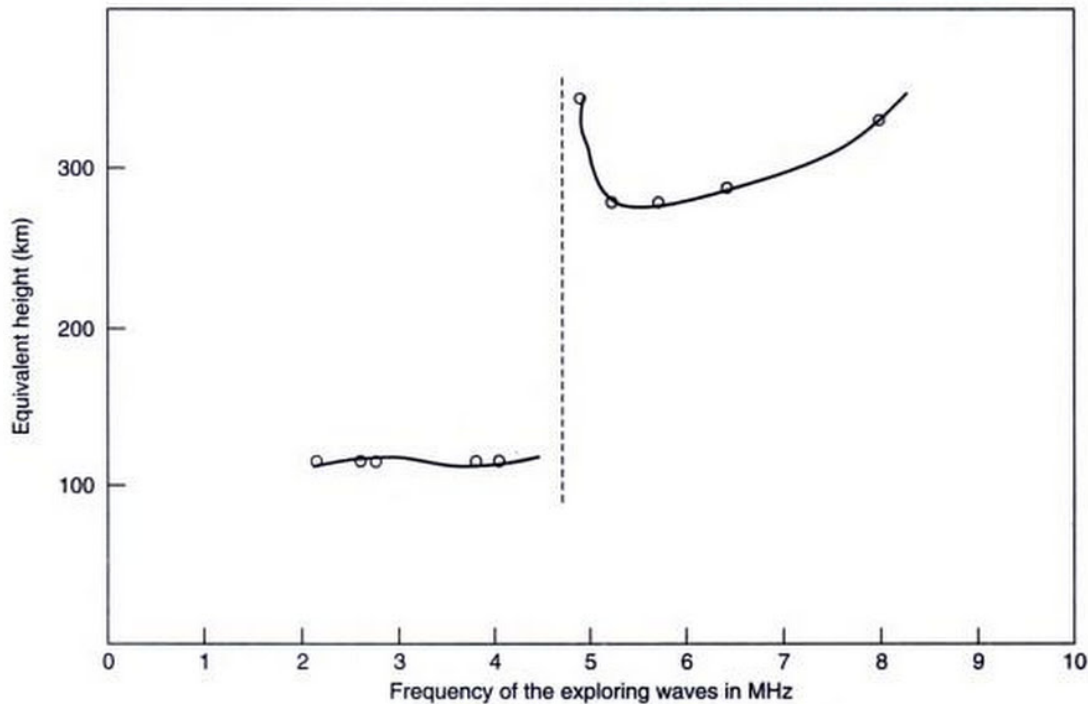


Fig. 2.21 Frequency dependence of the equivalent height of reflection from the *E* and *F* regions of the ionosphere. [Adapted from Ref. 9]

the most outstanding texts on the subject by Professor S. K. Mitra⁹.

Example 2.10 In this example we will obtain the solution of the ray equation for the refractive index variation given by

$$n^2(x) = n_0^2 + n_2^2 (1 - e^{-\alpha x}) \quad (68)$$

Substituting in Eq. (27) we would obtain

$$\begin{aligned} \pm dz &= \frac{\tilde{\beta} dx}{\left[(n_0^2 + n_2^2 - \tilde{\beta}^2) - n_2^2 e^{-\alpha x}\right]^{1/2}} \\ &= \frac{\tilde{\beta} e^{\alpha x/2} dx}{n_2 [K^2 e^{\alpha x} - 1]^{1/2}} \end{aligned}$$

or
$$\pm dz = \frac{2\tilde{\beta}}{K \alpha n_2} \frac{d\Phi}{(\Phi^2 - 1)^{1/2}} \quad (69)$$

where
$$K = \frac{1}{n_2} (n_0^2 + n_2^2 - \tilde{\beta}^2)^{1/2} \quad (70)$$

and
$$\Phi(x) = K e^{\alpha x/2} \quad (71)$$

The + and - sign in Eq. (69) correspond to a ray going up and a ray going down respectively. Further,

$$\tilde{\beta} = n_1 \cos \theta_1 \quad (72)$$

where θ_1 is the angle that the ray initially makes with the z -axis at $x = x_1$, $z = 0$ and $n_1 = n(x_1)$. Carrying out the elementary integration, we get

$$x(z) = \frac{2}{\alpha} \ln \left[\frac{1}{K} \cosh \gamma(C \pm z) \right] \quad (73)$$

where
$$\gamma = \frac{\alpha K n_2}{2\tilde{\beta}} \quad (74)$$

which gives us the ray path. Since $x = x_1$ at $z = 0$ (the initial point)

$$C = \frac{1}{\gamma} \cosh^{-1} (K e^{\alpha x_1/2}) \quad (75)$$

Further,

$$K e^{\alpha x_1/2} = \left[\frac{n_0^2 + n_2^2 - \tilde{\beta}^2}{n_0^2 + n_2^2 - n_1^2} \right]^{1/2} \quad (76)$$

Thus for a ray launched horizontally at $x = x_1$, $C = 0$. Typically ray paths (for different values of θ_1) are shown in Figs. 2.12 and 2.13.

*A proof for the applicability of Fermat's principle in anisotropic media has been given by Newcomb (Ref. 10); the proof, however, is quite complicated. Ray paths in biaxial media are discussed in Ref. 11.

2.5 REFRACTION OF RAYS AT THE INTERFACE BETWEEN AN ISOTROPIC MEDIUM AND AN ANISOTROPIC MEDIUM

In this section we will use Fermat's principle to determine the direction of the refracted ray for a ray incident at the interface of an isotropic and an anisotropic medium*. We may point out that in an isotropic medium the properties remain the same in all directions; typical examples are glass, water and air. On the other hand, in an anisotropic medium, some of the properties (such as speed of light) may be different in different directions. In Chapter 19 we will consider anisotropic media in greater detail; we may mention here that when a light ray is incident on a crystal like calcite, it (in general) splits into two rays known as ordinary and extraordinary rays. The velocity of the ordinary ray obeys Snell's laws but the extraordinary ray does not. We will now use Fermat's principle to study the refraction of a ray when it is incident from an isotropic medium into an anisotropic medium—both media are assumed to be homogeneous.

In a uniaxial medium, the refractive index variation for the extraordinary ray is given by [see Eq. (121) of Chapter 19]

$$n^2(\theta) = n_o^2 \cos^2 \theta + n_e^2 \sin^2 \theta \quad (77)$$

where n_o and n_e are constants of the crystal and θ represents the angle that the ray makes with the optic axis. Obviously, when the extraordinary ray propagates parallel to the optic axis (i.e., when $\theta = 0$), its speed is c/n_o and when it propagates perpendicular to the optic axis ($\theta = \pi/2$) its speed is c/n_e .

2.5.1 Optic Axis Normal to the Surface

We first consider the particularly simple case of the optic axis being normal to the surface. Referring to Fig. 2.22, the optical path length from A and B is given by

$$L_{op} = n_1 [h_1^2 + (L - x)^2]^{1/2} + n(\theta) [h_2^2 + x^2]^{1/2} \quad (78)$$

where n_1 is the refractive index of medium I and we have assumed the incident ray, the refracted ray and the optic axis to lie in the same plane. Since

$$\cos \theta = \frac{h_2}{(h_2^2 + x^2)^{1/2}} \quad \text{and} \quad \sin \theta = \frac{x}{(h_2^2 + x^2)^{1/2}}$$

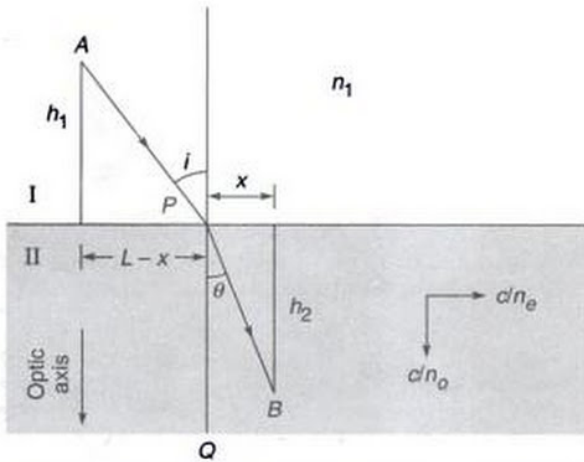


Fig. 2.22 The direction of the refracted extraordinary ray when the optic axis (of the uniaxial crystal) is normal to the surface.

we have

$$L_{op} = n_1 [h_1^2 + (L-x)^2]^{1/2} + [n_o^2 h_2^2 + n_e^2 x^2]^{1/2} \quad (79)$$

For the actual ray path, we must have

$$\frac{dL_{op}}{dx} = 0$$

implying

$$\frac{n_1 (L-x)}{[h_1^2 + (L-x)^2]^{1/2}} = \frac{n_e^2 x}{[n_o^2 h_2^2 + n_e^2 x^2]^{1/2}}$$

or

$$n_1 \sin i = \frac{n_e^2 \tan r}{[n_o^2 + n_e^2 \tan^2 r]^{1/2}} \quad (80)$$

where we have used the fact that the

$$\text{angle of refraction } r = \theta \quad \text{and} \quad \tan r = \frac{x}{h_2}.$$

Simple manipulations give us

$$\tan r = \frac{n_o n_1 \sin i}{n_e \sqrt{n_e^2 - n_1^2 \sin^2 i}} \quad (81)$$

using which we can calculate the angle of refraction for a given angle of incidence (when the optic axis is normal to the surface). As a simple example, we assume the first medium to be air so that $n_1 = 1$. Then

$$\tan r = \frac{n_o \sin i}{n_e \sqrt{n_e^2 - \sin^2 i}} \quad (\text{when } n_1 = 1) \quad (82)$$

If we assume the second medium to be calcite, then

$$n_o = 1.65836, \quad \text{and} \quad n_e = 1.48641$$

Thus for $i = 45^\circ$, we readily get

$$r \approx 31.1^\circ$$

It may be seen that if $n_o = n_e = n_2$ (say) then Eq. (80) simplifies to

$$n_1 \sin i = n_2 \sin r \quad (83)$$

which is nothing but Snell's law.

2.5.2 Optic Axis in the Plane of Incidence*

We next consider a more general case of the optic axis making an angle ϕ with the normal; however, the optic axis is assumed to lie in the plane of incidence as shown in Fig. 2.23. We may mention here that in general, in an anisotropic medium, the refracted ray does not lie in the plane of incidence. However, it can be shown that if the optic axis lies in the plane of incidence then the refracted ray also lies in the plane of incidence. In the present calculation, we are assuming this and finding the direction of the refracted ray

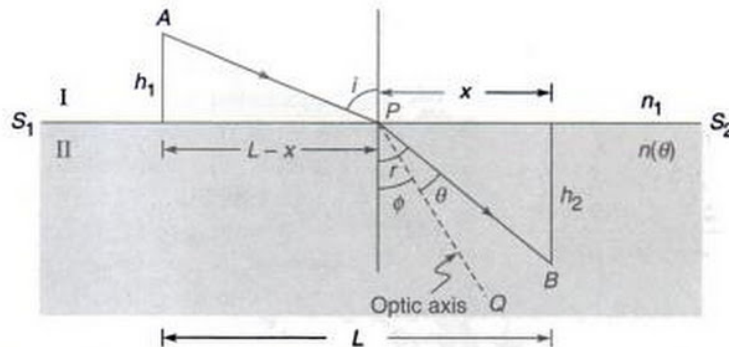


Fig. 2.23 The direction of the refracted extra-ordinary ray when the optic axis (of the uniaxial crystal) lies in the plane of incidence making an angle ϕ with the normal to the interface.

*May be skipped in the first reading.

for a given angle of incidence. Now, the optical path length from A to B (see Fig. 2.23) is given by

$$L_{\text{op}} = n_1[h_1^2 + (L-x)^2]^{1/2} + n(\theta)[h_2^2 + x^2]^{1/2} \quad (84)$$

Since $\theta = r - \phi$, we have

$$\begin{aligned} n^2(\theta) &= n_o^2 \cos^2(r - \phi) + n_e^2 \sin^2(r - \phi) \\ &= n_o^2 (\cos r \cos \phi + \sin r \sin \phi)^2 + \\ &\quad n_e^2 (\sin r \cos \phi - \cos r \sin \phi)^2 \end{aligned}$$

$$\begin{aligned} &= n_o^2 \left[\frac{h_2}{\sqrt{h_2^2 + x^2}} \cos \phi + \frac{x}{\sqrt{h_2^2 + x^2}} \sin \phi \right]^2 + \\ &\quad n_e^2 \left[\frac{x}{\sqrt{h_2^2 + x^2}} \cos \phi - \frac{h_2}{\sqrt{h_2^2 + x^2}} \sin \phi \right]^2 \end{aligned}$$

Thus

$$\begin{aligned} n(\theta) &= \frac{1}{\sqrt{h_2^2 + x^2}} [n_o^2 (h_2 \cos \phi + x \sin \phi)^2 \\ &\quad + n_e^2 (x \cos \phi - h_2 \sin \phi)^2]^{1/2} \quad (85) \end{aligned}$$

and

$$\begin{aligned} L_{\text{op}} &= n_1[h_1^2 + (L-x)^2]^{1/2} + \\ &\quad [n_o^2 (h_2 \cos \phi + x \sin \phi)^2 + n_e^2 (x \cos \phi - h_2 \sin \phi)^2]^{1/2} \quad (86) \end{aligned}$$

For the actual ray path, we must have

$$\frac{dL_{\text{op}}}{dx} = 0$$

implying

$$\begin{aligned} \frac{n_1(L-x)}{[h_1^2 + (L-x)^2]^{1/2}} &= \\ \frac{n_o^2 (h_2 \cos \phi + x \sin \phi) \sin \phi + n_e^2 (x \cos \phi - h_2 \sin \phi) \cos \phi}{[n_o^2 (h_2 \cos \phi + x \sin \phi)^2 + n_e^2 (x \cos \phi - h_2 \sin \phi)^2]^{1/2}} \\ \text{or } n_1 \sin i &= \frac{n_o^2 \cos \theta \sin \phi + n_e^2 \sin \theta \cos \phi}{[n_o^2 \cos^2 \theta + n_e^2 \sin^2 \theta]^{1/2}} \quad (87) \end{aligned}$$

For given values of the angles i and ϕ , the above equation can be solved to give the values of θ and hence the angle of refraction $r (= \theta + \phi)$.

Some interesting particular cases may be noted.

- (i) When $n_o = n_e = n_2$, the anisotropic medium becomes isotropic and Eq. (87) simplifies to

$$n_1 \sin i = n_2 \sin(\theta + \phi) = n_2 \sin r$$

which is nothing but Snell's law.

- (ii) When $\phi = 0$, i.e., the optic axis is normal to the surface, Eq. (87) becomes

$$\begin{aligned} n_1 \sin i &= \frac{n_e^2 \sin \theta}{[n_o^2 \cos^2 \theta + n_e^2 \sin^2 \theta]^{1/2}} \\ &= \frac{n_e^2 \sin r}{[n_o^2 \cos^2 r + n_e^2 \sin^2 r]^{1/2}} \quad (88) \end{aligned}$$

where we have used the fact that $r = \theta$. The above equation is identical to Eq. (80).

- (iii) Finally, we consider normal incidence, i.e., $i = 0$. Thus, Eq. (87) gives us

$$n_o^2 \cos \theta \sin \phi + n_e^2 \sin \theta \cos \phi = 0$$

or

$$n_o^2 \cos(r - \phi) \sin \phi + n_e^2 \sin(r - \phi) \cos \phi = 0$$

or

$$\begin{aligned} \cos r [n_o^2 \cos \phi \sin \phi - n_e^2 \sin \phi \cos \phi] \\ + \sin r [n_o^2 \sin^2 \phi + n_e^2 \cos^2 \phi] = 0 \end{aligned}$$

or

$$\tan r = \frac{(n_e^2 - n_o^2) \sin \phi \cos \phi}{n_o^2 \sin^2 \phi + n_e^2 \cos^2 \phi} \quad (89)$$

Equation (89) shows that in general $r \neq 0$ (see Fig. 2.24). We may mention here that, for normal incidence, the above analysis is valid for an arbitrary orientation of the optic axis; the refracted (extra-ordinary) ray lies in the plane containing the normal and the optic axis. Furthermore, for normal incidence, when the crystal is rotated about the normal, the refracted ray also rotates on the surface of a cone [see Fig. 19.16(b)].

Returning to Eq. (89) we note that when the optic axis is normal to the surface ($\phi = 0$) or when the optic axis is parallel to the surface but lying in the plane of incidence ($\phi = \pi/2$), $r = 0$ and the ray goes undeviated.

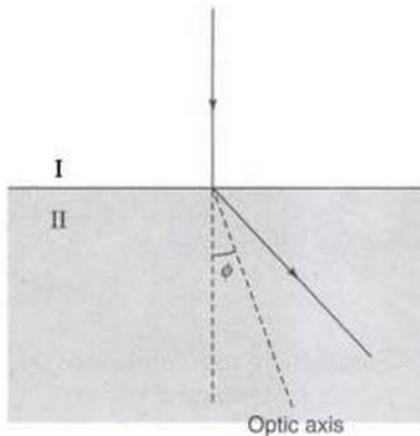


Fig. 2.24 For normal incidence, in general, the refracted extraordinary ray undergoes finite deviation. However, the ray proceeds undeviated when the optic axis is parallel or normal to the surface.

SUMMARY

- The slightly modified version of Fermat's principle is: *the actual ray path between two points is the one for which the optical path length is stationary with respect to variations of the path.*
- Laws of reflections and Snell's law of refraction ($n_1 \sin \phi_1 = n_2 \sin \phi_2$, where ϕ_1 and ϕ_2 represent the angles of incidence and refraction) can be derived from Fermat's principle.
- For an inhomogeneous medium characterized by the refractive index variation $n(x)$, the ray paths $[x(z)]$ are such that the product $n(x) \cos \theta(x)$ remains constant, here $\theta(x)$ is the angle that the ray makes with the z -axis; this constant is denoted by $\tilde{\beta}$ which is known as the ray invariant. The exact ray paths are determined by solving either of the equations:

$$\frac{dx}{dz} = \pm \frac{\sqrt{n^2(x) - \tilde{\beta}^2}}{\tilde{\beta}}$$

or

$$\frac{d^2x}{dz^2} = \frac{1}{2\tilde{\beta}^2} \frac{dn^2(x)}{dx}$$

where the invariant $\tilde{\beta}$ is determined from the initial launching condition of the ray.

- Ray paths obtained by solving the ray equation can be used to study mirage, looming and also reflections from the ionosphere.
- In a parabolic index medium $n^2(x) = n_1^2 - \gamma^2 x^2$, the ray paths are sinusoidal:

$$x(z) = \pm x_0 \sin \Gamma z$$

where $\Gamma = \frac{1}{\tilde{\beta}}$, $x_0 = \frac{1}{\gamma} \sqrt{n_1^2 - \tilde{\beta}^2}$ and we have assumed

$z = 0$ where $x = 0$. Rays launched at different angles take approximately the same time in propagating through a large length of the medium.

- Fermat's principle can be used to study refraction of rays at the interface of an isotropic medium and an anisotropic medium.

PROBLEMS

- 2.1** In this and the following two problems we will use Fermat's principle to derive laws governing paraxial image formation by spherical mirrors.

Consider an object point O in front of a concave mirror whose center of curvature is at the point C . Consider an arbitrary point Q on the axis of the system and using a method similar to that used in Example 2.3, show that the optical path length L_{op} ($= OS + SQ$) is approximately given by

$$L_{op} \approx x + y + \frac{1}{2} r^2 \left[\frac{1}{x} + \frac{1}{y} - \frac{2}{r} \right] \theta^2 \quad (90)$$

where the distances x , y and r and the angle θ are defined in Fig. 2.25; θ is assumed to be small. Determine the paraxial image point and show that the result is consistent with the mirror equation

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R} \quad (91)$$

where u and v are the object and image distance and R is the radius of curvature with the sign convention that all distances to the right of P are positive and to its left negative.

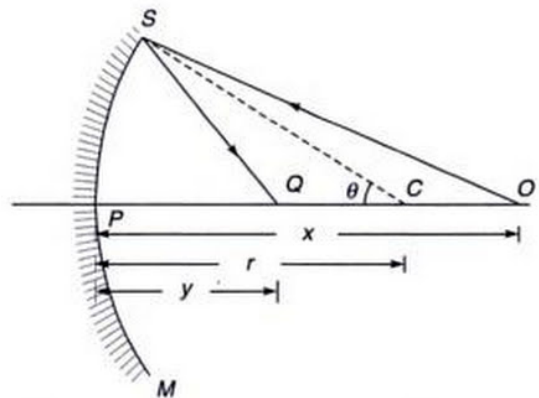


Fig. 2.25 Paraxial image formation by a concave mirror.

2.2 Fermat's principle can also be used to determine the paraxial image points when the object forms a virtual image. Consider an object point O in front of the convex mirror SPM (see Fig. 2.26). One should now assume the optical path length L_{op} to be $OS - SQ$; the minus sign occurs because the rays at S point away from Q (see Example 2.4). Show that

$$L_{op} \approx OS - SQ = x - y + \frac{1}{2} r^2 \left[\frac{1}{x} - \frac{1}{y} + \frac{2}{r} \right] \theta^2 \quad (92)$$

where the distances x , y and r and the angle θ are defined in Fig. 2.26. Show that the paraxial image is formed at $y = y_0$ which is given by

$$\frac{1}{x} - \frac{1}{y_0} = -\frac{2}{r} \quad (93)$$

which is consistent with Eq. (91) because whereas the object distance u is positive, the image distance v and the radius of curvature R are negative since the image point and the center of curvature lie on the left of the point P .

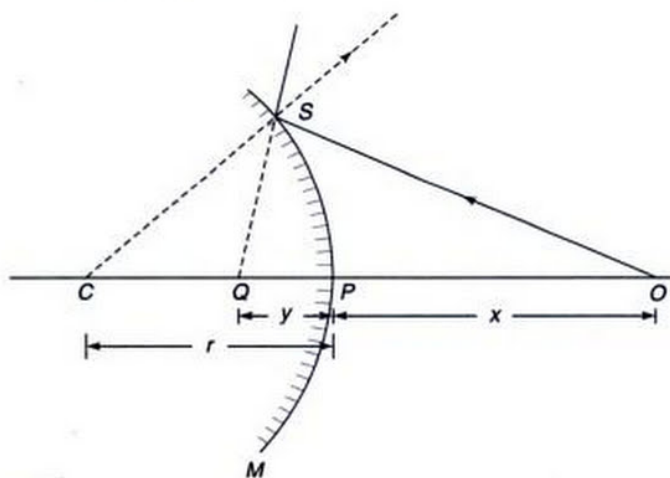


Fig. 2.26 Paraxial image formation by a convex mirror.

2.3 Proceeding as in the previous problem, use Fermat's principle to determine the mirror equation for an object point at a distance less than $R/2$ from a concave mirror of radius of curvature R .

2.4 We next consider a point object O in front of a concave refracting surface SPM separating two media of refracting indices n_1 and n_2 (see Fig. 2.27); C represents the center of curvature. In this case also one obtains a virtual image. Let Q represent an arbitrary point on the axis. We now have to consider the optical path length $L_{op} = n_1 OS - n_2 SQ$; show that it is given by

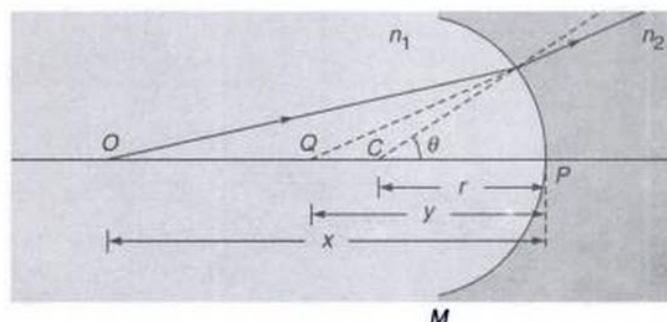


Fig. 2.27 Paraxial image formation by a concave refracting surface SPM.

$$\begin{aligned} L_{op} &= n_1 OS - n_2 SQ \\ &\approx n_1 x - n_2 y - \frac{1}{2} r^2 \left[\frac{n_2}{y} - \frac{n_1}{x} - \frac{n_2 - n_1}{r} \right] \theta^2 \end{aligned} \quad (94)$$

Also show that the above expression leads to the paraxial image point which is consistent with Eq. (10); we may note that u , v and R are all negative quantities because they are on the left of the refracting surface.

2.5 If we rotate an ellipse about its major axis we obtain what is known as an ellipsoid of revolution. Show by using Fermat's principle that all rays parallel to the major axis of the ellipse will focus to one of the focal points of the ellipse (see Fig. 2.28), provided the eccentricity of the ellipse equals n_1/n_2 . (Hint: Start with the condition that

$$n_2 AC' = n_1 QB + n_2 BC$$

and show that the point B (whose coordinates are x and y) lies on the periphery of an ellipse).

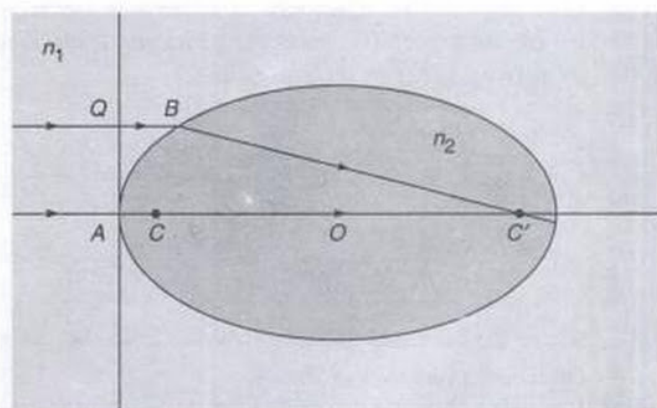


Fig. 2.28 All rays parallel to the major axis of the ellipsoid of revolution will focus to one of the focal points of the ellipse provided the eccentricity $= n_1/n_2$.

- 2.6 C is the center of the reflecting sphere of radius R (see Fig. 2.29). P_1 and P_2 are two points on a diameter equidistant from the center. (a) Obtain the optical path length $P_1O + OP_2$ as a function of θ , and (b) find the values of θ for which P_1OP_2 is a ray path from reflection at the sphere.

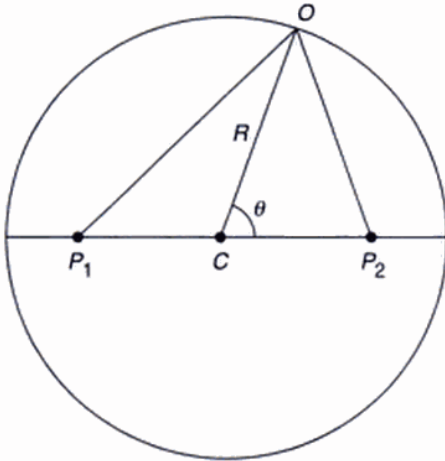


Fig. 2.29 A spherical reflector.

- 7 SPM is a spherical refracting surface separating two media of refractive indices n_1 and n_2 . (see Fig. 2.30). Consider an object point O forming a virtual image at the point I . We assume that all rays emanating from O appear to emanate from I so as to form a perfect image. Thus according to Fermat's principle, we must have

$$n_1 OS - n_2 SI = n_1 OP - n_2 PI$$

where S is an arbitrary point on the refracting surface. Assuming the right hand side to be zero, show that the refracting surface is spherical, with the radius given by

$$r = \frac{n_1}{n_1 + n_2} OP \quad (95)$$

Thus show that

$$n_1^2 d_1 = n_2^2 d_2 = n_1 n_2 r \quad (96)$$

where d_1 and d_2 are defined in Fig. 2.30. (see also Fig. 3.12 and Sec. 3.10).

[Hint: We consider a point C which is at a distance d_1 from the point O and d_2 from the point I . Assume the origin to be at O and let (x, y, z) represent the coordinates of the point S . Thus

$$\begin{aligned} n_1 (x^2 + y^2 + z^2)^{1/2} - n_2 (x^2 + y^2 + \Delta^2)^{1/2} \\ = n_1 (r + d_1) - n_2 (r + d_2) = 0 \end{aligned}$$

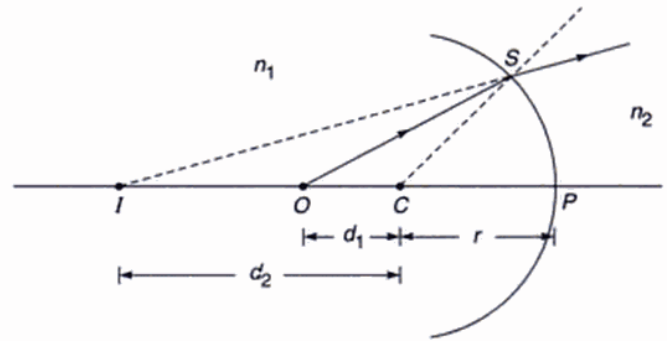


Fig. 2.30 All rays emanating from O and getting refracted by the spherical surface SPM appear to come from I .

where $\Delta = d_2 - d_1$. The above equation would give the equation of a sphere whose center is at a distance of $n_2 r / n_1 (= d_1)$ from O .]

- 2.8 Referring to Fig. 2.31, if I represents a perfect image of the point O , show that the equation of the refracting surface (separating two media of refractive indices n_1 and n_2) is given by

$$\begin{aligned} n_1 [x^2 + y^2 + z^2]^{1/2} + n_2 [x^2 + y^2 + (z_2 - z)^2]^{1/2} \\ = n_1 z_1 - n_2 (z_2 - z_1) \end{aligned} \quad (97)$$

where the origin is assumed to be at the point O and the coordinates of P and I are assumed to be $(0, 0, z_1)$ and $(0, 0, z_2)$ respectively. The surface corresponding to Eq. (97) is known as a Cartesian oval.

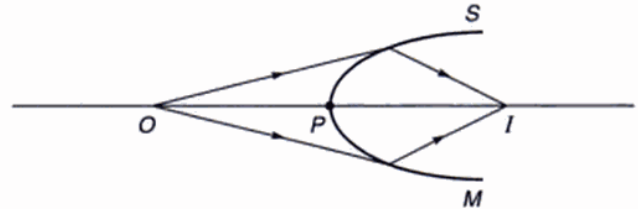


Fig. 2.31 The Cartesian oval. All rays emanating from O and getting refracted by SPM pass through I .

- 2.9 For the refractive index variation given by Eqs. (21) and (22), a ray is launched at $x = 0.43$ m making an angle $-\pi/60$ with the z -axis (see Fig. 2.12). Calculate the value of x at which it will become horizontal.

- 2.10 For the refractive index variation given by Eqs. (21) and (22), a ray is launched at $x = 2.8$ m such that it becomes horizontal at $x = 0.2$ m (see Fig. 2.13). Calculate the angle that the ray will make with the z -axis at the launching point.

[Ans: $\theta_1 \approx 19^\circ$]

- 2.11** Consider a parabolic index medium characterized by the following refractive index variation:

$$n^2(x) = n_1^2 \left[1 - 2\Delta \left(\frac{x}{a} \right)^2 \right] \quad |x| < a$$

$$= n_2^2 \quad |x| > a$$

Assume $n_1 = 1.50$, $n_2 = 1.48$, $a = 50 \mu\text{m}$. Calculate the value of Δ .

- (a) Assume rays launched on the axis at $z = 0$ (i.e., $x = 0$ when $z = 0$) with

$$\tilde{\beta} = 1.495, 1.490, 1.485, 1.480, 1.475 \text{ and } 1.470$$

In each case calculate the angle that the ray initially makes with the z -axis (θ_1) and plot the ray paths. In each case find the height at which the ray becomes horizontal.

- (b) Assume rays incident normally on the plane $z = 0$ at $x = 0, \pm 10 \mu\text{m}, \pm 20 \mu\text{m}, \pm 30 \mu\text{m}, \pm 40 \mu\text{m}$.

Find the corresponding values of $\tilde{\beta}$, calculate the focal length for each ray and qualitatively plot the ray paths.

- 2.12** In an inhomogeneous medium the refractive index is given by

$$n^2(x) = 1 + \frac{x}{L} \quad \text{for } x > 0$$

$$= 1 \quad \text{for } x < 0$$

Write down the equation of a ray (in the x - z plane) passing through the point $(0, 0, 0)$ where its orientation with respect to x axis is 45° .

- 2.13** For the refractive index profile given by Eq. (23), show that Eq. (27) can be written in the form

$$\pm \frac{\alpha K_1 n_2}{2\tilde{\beta}} dz = \frac{dG}{\sqrt{1 - G^2}} \quad (98)$$

where

$$K_1 = \frac{\sqrt{\tilde{\beta}^2 - n_0^2}}{n_2} \quad \text{and} \quad G(x) = K_1 e^{\alpha x/2} \quad (99)$$

Integrate Eq. (98) to determine the ray paths.

- 2.14** Consider a graded index medium characterised by the following refractive index distribution

$$n^2(x) = n_1^2 \operatorname{sech}^2 gx \quad (100)$$

Substitute in Eq. (27) and integrate to obtain

$$x(z) = \frac{1}{g} \sinh^{-1} \left[\frac{\sqrt{n_1^2 - \tilde{\beta}^2}}{\tilde{\beta}} \sin gz \right] \quad (101)$$

Notice that the periodic length

$$z_p = \frac{2\pi}{g}$$

is independent of the launching angle (see Fig. 2.32) and all rays rigorously take the same amount of time in propagating through a distance z_p in the z -direction.

[Hint: While carrying out the integration, make the

$$\text{substitution: } \zeta = \frac{\tilde{\beta}}{\sqrt{n_1^2 - \tilde{\beta}^2}} \sinh gx]$$

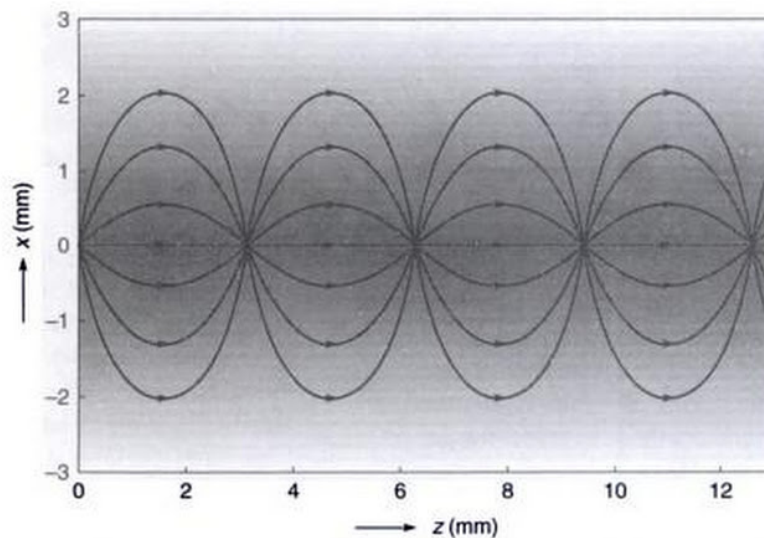


Fig. 2.32 Ray paths in a graded index medium characterized by Eq. (100).

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, UK, 1975.
2. A.K. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, (1978). [Reprinted by Macmillan India, New Delhi].
3. R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1965.
4. M.S. Sodha, A.K. Aggarwal and P.K. Kaw, 'Image formation by an optically stratified medium: Optics of mirage and looming', *British Journal of Applied Physics*, vol. **18**, 503, 1967.
5. R.T. Bush and R.S. Robinson, 'A Note explaining the mirage', *American Journal of Physics*, vol. **42**, 774, 1974.
6. A.B. Fraser and W.H. Mach, 'Mirages', *Scientific American*, January, vol. **234**, 102, 1976.
7. E. Khular, K. Thyagarajan and A. K. Ghatak, 'A note on mirage formation', *American Journal of Physics*, vol. **45**, 90, 1977.
8. W.J. Humphreys, *Physics of the Air*, McGraw-Hill Book Co., New York, 1920.
9. S. K. Mitra, *The Upper Atmosphere*, Second Edition, The Asiatic Society, Calcutta, 1952.
10. W.A. Newcomb, 'Generalized Fermat's principles', *American Journal of Physics*, vol. **51**, 338, 1983.
11. E. Khular, K. Thyagarajan and A.K. Ghatak, 'Ray tracing in uniaxial and biaxial media', *Optik*, vol. **46**, 297, 1976.
12. V Lakshminarayanan, A. Ghatak and K Thyagarajan, *Lagrangian Optics*, Kluwer Academic Publishers, 2002.

Chapter 3

Refraction and Reflection by Spherical Surfaces

The uses of plane and curved mirrors and of convex and concave lenses were discovered independently in China and in Greece. References to burning mirrors go back almost to the start of history, and it is possible that Chinese and Greek knowledge were both derived from a common source in Mesopotamia, India or Egypt... Pythagoras, Greek philosopher and mathematician (6th century BC), suggested that light consists of rays that, acting like feelers, travel in straight lines from the eye to the object and that the sensation of sight is obtained when these rays touch the object. In this way, the more mysterious sense of sight is explained in terms of the intuitively accepted sense of touch. It is only necessary to reverse the direction of these rays to obtain the basic scheme of modern geometrical optics. The Greek mathematician Euclid (300 BC), who accepted the Pythagorean idea, knew that the angle of reflected light rays from a mirror equals the angle of incident light rays from the object to the mirror. The idea that light is emitted by a source and reflected by an object and then enters the eye to produce the sensation of sight was known to Epicurus, another Greek philosopher (300 BC). The Pythagorean hypothesis was eventually abandoned and the concept of rays traveling from the object to the eye was finally accepted about AD 1000 under the influence of an Arabian mathematician and physicist named Alhazen.

—The New Encyclopedia Britannica, Vol. 23

Alhazen had used spherical and parabolic mirrors and was aware of spherical aberration. He also investigated the magnification produced by lenses and atmospheric refraction. His work was translated into Latin and became accessible to later European scholars.

—From the Internet

3.1 INTRODUCTION

In this chapter we will study the formation of an image by simple optical systems. We will assume the optical system to be made up of a number of refracting surfaces like a combination of lenses.* In order to trace a ray through such an optical system, it is necessary only to apply Snell's laws at each refracting surface which are as follows:

- (a) the incident ray, the refracted ray and the normal (to the surface) lie in the same plane; and
- (b) if ϕ_1 and ϕ_2 represent the angles of incidence and refraction respectively, then

$$\frac{\sin \phi_1}{\sin \phi_2} = \frac{n_2}{n_1} \quad (1)$$

where n_1 and n_2 are the refractive indices of the two media (see Fig. 3.1). Although there is no additional physics involved (other than the Snell's laws) in the tracing of rays, the design of even a simple optical system involves tracing many rays and therefore considerable numerical computations. Nowadays, such numerical computations are usually

done on a high-speed computer. It may be of interest to note that optical designers were among the first to make use of electronic computers when they were introduced in the early fifties.

3.2 REFRACTION AT A SINGLE SPHERICAL SURFACE

We will first consider refraction at a spherical surface *SPM* separating two media of refractive indices n_1 and n_2 (see Fig. 3.1(a)). Let C represent the center of curvature of the spherical surface. We will consider a point object O emitting rays in all directions. We will use Snell's laws of refraction to determine the image of the point O . We may mention that not all rays emanating from O converge to a single point; however, if we consider only those rays which make small angles with the line joining the points O and C then all rays do converge to a single point I [see Fig. 3.1(a)]. This is known as the *paraxial approximation* and according to Fermat's principle all paraxial rays take the same amount of time to travel from O to I (see Example 2.3).

*The optical system may also consist of mirrors, in which case the reflection of rays should also be taken into account (see Section 3.3).

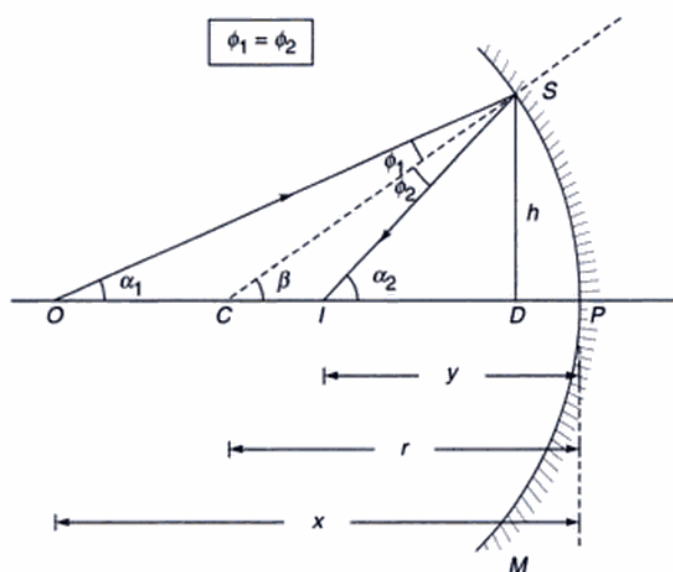


Fig. 3.3 Paraxial image formation by a spherical reflecting surface SPM.

If we again use the sign convention that all the distances to the right of P are positive and those to its left negative, then $u = -x$, $v = -y$ and $R = -r$; thus we obtain the mirror equation

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R} \quad (7)$$

which is the same as was derived by using Fermat's principle (see Problem 2.1). It is interesting to note that if we set $n_2 = -n_1$ in Eq. (5), we get Eq. (7). This follows from the fact that Snell's law of refraction (Eq. 1) becomes the law of reflection if we have $n_2 = -n_1$.

We illustrate the use of Eq. (7) through an example.

Example 3.2 Consider an optical system consisting of a concave mirror $S_1P_1M_1$ and convex mirror $S_2P_2M_2$ of radii of curvatures 60 and 20 cm respectively (see Fig. 3.4). We would like to determine the final image position of the object point O which is at a distance of 80 cm from the point P_1 , the two mirrors being separated by a distance of 40 cm.

We first consider the imaging by $S_1P_1M_1$; since $u = -80$ cm and $R = -60$ cm (because both O and C are on the left of P_1), we have

$$-\frac{1}{80} + \frac{1}{v} = -\frac{2}{60} \Rightarrow v = -48 \text{ cm}$$

In the absence of the mirror $S_2P_2M_2$, a real image will be formed at I_1 which now acts as a virtual object for $S_2P_2M_2$. Since I_1 is to the left of P_2 , we have (considering imaging by $S_2P_2M_2$), $u = -8$ cm and $R = -20$ cm, giving

$$\frac{1}{v} - \frac{1}{8} = -\frac{2}{20} \Rightarrow v = +40 \text{ cm}$$

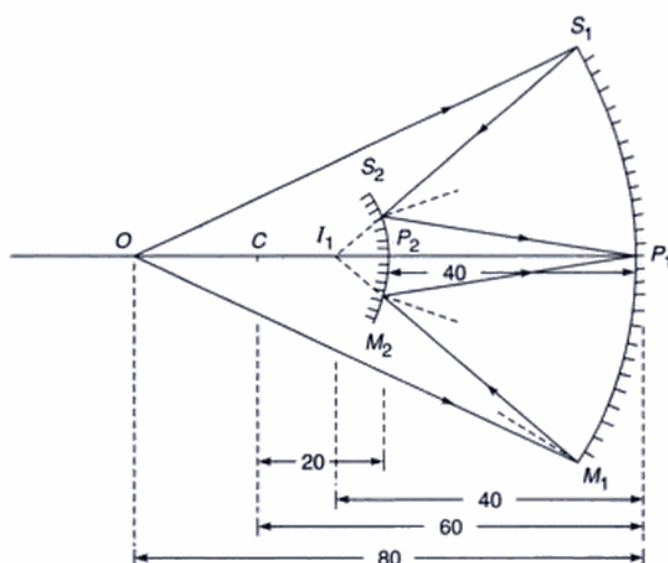


Fig. 3.4 Paraxial image formation by an optical system comprising a concave mirror $S_1P_1M_1$ and a convex mirror $S_2P_2M_2$.

Thus the final image is formed on the right of $S_2P_2M_2$ at a distance of 40 cm, which happens to be the point P_1 .

3.4 THE THIN LENS

A medium bounded by two spherical refracting surfaces is referred to as a spherical lens. If the thickness of such a lens (shown as t in Fig. 3.5) is very small compared to object and image distances and to the radii of curvature of the refracting surfaces then the lens is referred to as a thin spherical lens. In general, a lens may have non-spherical refracting surfaces (e.g. it may have cylindrical surfaces). However, most lenses employed in optical systems have spherical refracting surfaces. Therefore, we will simply use the term 'lens' to imply a spherical lens. Different types of lenses are shown in Fig. 3.6. The line joining the centers of curvature of the spherical refracting surfaces is referred to as the *axis* of the lens.

In this section, we will consider the paraxial image formation by a thin lens. The corresponding considerations for a thick lens will be discussed in Problem 3.6.

We consider a point object O on the axis of a (thin) lens as shown in Fig. 3.5. The lens is placed in a medium of refractive index n_1 and the refractive index of the material of the lens is n_2 . Let R_1 and R_2 be the radii of curvature of the left and right surfaces of the lens; for the lens shown in Fig. 3.5, R_1 is positive and R_2 is negative. In order to determine the position of the image we will consider successive refractions at the two surfaces; the image formed by the first surface is considered as the object (which may

$$\frac{1}{f_2} = \frac{1}{n_3} \left[\frac{n_2 - n_1}{R_1} + \frac{n_3 - n_2}{R_2} \right] \quad (16)$$

Once we know f_1 and f_2 (and therefore the positions of the first and second principal foci) the (paraxial) image can be graphically constructed from the following rules:

- (i) A ray passing through the first principal focus will, after refraction, emerge parallel to the axis [see ray 1 in Figs. 3.8(a) and (b)]
- (ii) A ray parallel to the axis will, after refraction, either pass through or appear to come from (depending on the sign of f_2) the second principal focus [see ray 2 in Figs. 3.8(a) and (b)]
- (iii) A ray passing through the center of the lens P will pass through undeviated* [see ray 3 in Figs 3.8(a) and (b)]

3.6 THE NEWTON FORMULA

Let x_1 be the distance of the object from the first principal focus F_1 (x_1 will be positive if the object point is on the right of F_1 and conversely) and let x_2 be the distance of the image from the second principal focus F_2 as shown in Figs 3.8(a) and (b). Considering similar triangles in Fig. 3.8(a), we have

$$\frac{-y'}{y} = \frac{-f_1}{-x_1} \quad (17)$$

and

$$\frac{-y'}{y} = \frac{x_2}{f_2} \quad (18)$$

where the vertical distances are positive if measured above the line and negative if measured below the line (see Sec. 3.2.1). Equations (17) and (18) give

$$f_1 f_2 = x_1 x_2 \quad (19)$$

which is known as the *Newtonian lens formula*. It may be noted that for a diverging lens [see Fig. 3.8(b)], Eqs. (17) and (18) would be

$$\frac{y'}{y} = \frac{f_1}{-x_1} = \frac{x_2}{-f_2}$$

which are identical to Eqs (17) and (18).

When the thin lens has the same medium on the two sides, then using Eq. (13) we have

$$x_1 x_2 = -f^2 \quad (20)$$

showing that x_1 and x_2 must be of opposite signs. Thus if the object lies on the left of the first principal focus, then the image will lie on the right of the second principal focus, and vice versa.

3.7 LATERAL MAGNIFICATION

The lateral magnification m is the ratio of the height of the image to that of the object. Considering either Fig. 3.8(a) or 3.8(b) we readily get

$$m = \frac{y'}{y} = \frac{v}{u} = \frac{f_2 + x_2}{f_1 + x_1} = -\frac{f_1}{x_1} = -\frac{x_2}{f_2} \quad (21)$$

where we have made use of Eqs (17) and (18). Obviously, if m is positive, the image is erect [as in Fig. 3.8(b) and conversely if m is negative, the image is inverted as in Fig. 3.8(a)].

The magnification can also be calculated as the product of the individual magnifications produced by each of the refracting surfaces; referring to Fig. 3.9, the magnification produced by a single refracting surface is given by

$$m = \frac{y'}{y}$$

and considering triangles AOC and ICB , we get

$$\frac{-y'}{y} = \frac{v - R}{-u + R} = \frac{\frac{v}{R} - 1}{-\frac{u}{R} + 1} \quad (22)$$

Now, Eq. (5) gives us

$$\frac{n_2}{n_1} - \frac{v}{u} = \frac{n_2 - n_1}{n_1} \frac{v}{R}$$

and

$$\frac{u}{v} - \frac{n_1}{n_2} = \frac{n_2 - n_1}{n_2} \frac{u}{R}$$

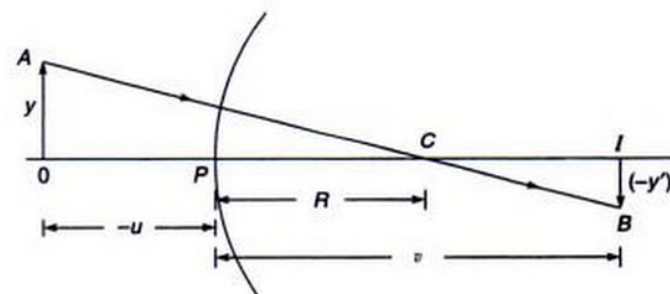


Fig. 3.9 Imaging of an object of height y by a spherical refracting surface.

*This follows from the fact that, for a thin lens, when $u = 0$, v is also equal to zero [see Eqs (10) and (14)].

Substituting for v/R and u/R in Eq. (22), we get

$$m = \frac{y'}{y} = \frac{n_1 v}{n_2 u} \quad (23)$$

Thus, if m_1 and m_2 represent the magnifications produced by the two refracting surfaces in Fig. 3.8, then

$$m = \frac{n_1}{n_2} \frac{v'}{u}$$

and

$$m_2 = \frac{n_2}{n_1} \frac{v}{v'}$$

where v' represents the distance of the image formed by the first refracting surface. Thus

$$m = m_1 m_2 = \frac{v}{u} \quad (24)$$

consistent with Eq. (21).

Example 3.3 Consider a system of two thin lenses as shown in Fig. 3.10. The convex lens has a focal length of +20 cm and the concave lens has a focal length of -10 cm. The two lenses are separated by 8 cm. For an object of height 1 cm (at a distance of 40 cm from the convex lens), calculate the position and size of the image. (The same problem will be solved again in Chapter 4 by using the matrix method.)

Solution: Let us first calculate the position and size of the image formed by the first lens:

$$u = -40 \text{ cm}, f = +20 \text{ cm}$$

Therefore, using Eq. (11) we get

$$\frac{1}{v} = \frac{1}{u} + \frac{1}{f} = -\frac{1}{40} + \frac{1}{20} = +\frac{1}{40}$$

Thus, $v = +40$ cm and $m_1 = -1$; the image is of the same size but inverted. This image acts as a virtual object for the concave lens with $u = +32$ cm and $f = -10$ cm. Thus

$$\frac{1}{v} = \frac{1}{32} - \frac{1}{10} = -\frac{22}{320}$$

giving

$$v \approx -14.5 \text{ cm}$$

Further,

$$m_2 = -\frac{320/22}{32} = -\frac{1}{2.2}$$

Thus

$$m = m_1 m_2 = +\frac{1}{2.2}$$

The final image is formed at a distance of 14.5 cm on the left of the concave lens. The image is virtual, erect and smaller in size by a factor of 2.2.

3.8 APLANATIC POINTS OF A SPHERE

In Section 3.2, while discussing image formation by a single refracting surface we had made use of the paraxial approximation, i.e., we had considered rays which made small angles with the axis. In this approximation, it was found that the images of point objects are perfect, i.e., *all* rays emanating from a given object point were found to intersect at *one* point which is the image point. If we had considered rays which make large angles with the axis, then we would have observed that, in general, (after refraction) they do not pass through the same point on the axis (see Fig. 3.11) and a perfect image is not formed. The image is said to be afflicted with aberrations. However, for a given spherical surface, there exist two points for which *all* rays

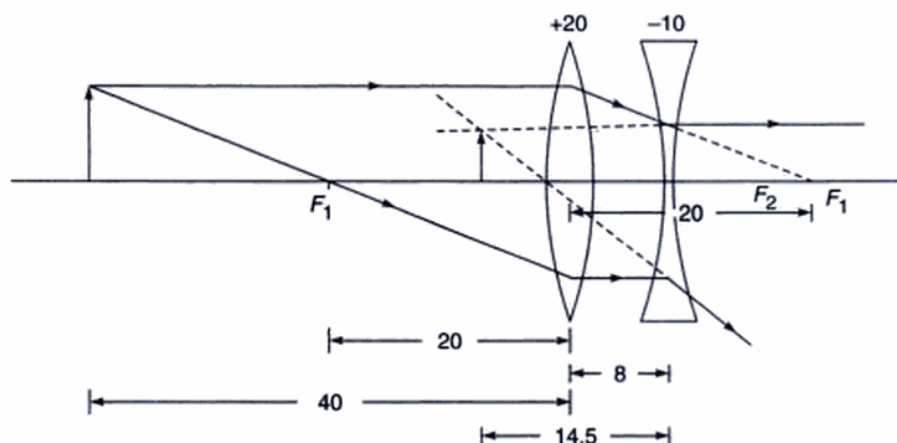


Fig. 3.10 Paraxial imaging by an optical system consisting of a converging lens of focal length 20 cm and a diverging lens of focal length -10 cm separated by 8 cm. All distances in the figure are in centimeters.

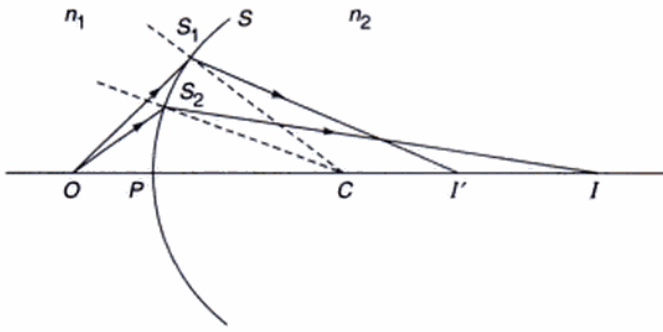


Fig. 3.11 The point I represents the paraxial image point of the object point O formed by a spherical refracting surface SPM . However, if we consider non-paraxial rays like OS_1 (which make large angles with the axis) then the refracted ray, in general, will not pass through the point I —this leads to aberrations in the image.

emanating from one point intersect each other at the other point. This point is at a distance equal to $n_2 |R|/n_1$ from the center of the spherical surface and a virtual image is formed at a distance of $n_1 |R|/n_2$ from the center [see Figs 3.12(a) and (b)]. This can be easily proved by using Fermat's principle (see Problem 2.7) or by using geometrical methods (see Sec. 3.10). The two points are said to be the *aplanatic points* of the sphere and are utilized in the construction of aplanatic lenses (see Fig. 3.13) which are used in wide aperture oil immersion microscope objectives. The points O and I are the aplanatic points of the spherical surface of radius R_2 (see Fig. 3.13). Thus

$$OP_2 = |R_2| \left[1 + \frac{n_1}{n_2} \right] \quad (25)$$

and

$$IP_2 = |R_2| \left[1 + \frac{n_2}{n_1} \right] \quad (26)$$

Now, the radius of curvature of the first surface ($= R_1$) is such that the point O coincides with its center of curvature. Hence *all* rays emanating from O hit the first surface normally and move on undeviated. Therefore, for all practical purposes, we may assume O to be embedded in a medium of refractive index n_2 . A perfect (virtual) image of O is formed at I .

3.8.1 The Oil Immersion Objective

The principle of aplanatism has a very important application in microscope objectives where one is interested in having as wide a pencil of light as possible without causing any aberrations. We refer to the optical system shown in Fig. 3.14. The hemispherical lens L_1 is placed in contact with a drop of oil whose refractive index is the same as that of the lens. The object O is immersed in the oil and the distance OC is made equal to $n_3 |R_1|/n_2$ so that the point O is the aplanatic point with respect to the hemispherical surface, which is why a perfect (virtual) image is formed at I_1 . Now L_2 is an aplanatic lens with respect to the object point at I_1 and therefore a perfect image of I_1 is formed at I . The lateral magnifications caused by the refracting surface R_1 and lens

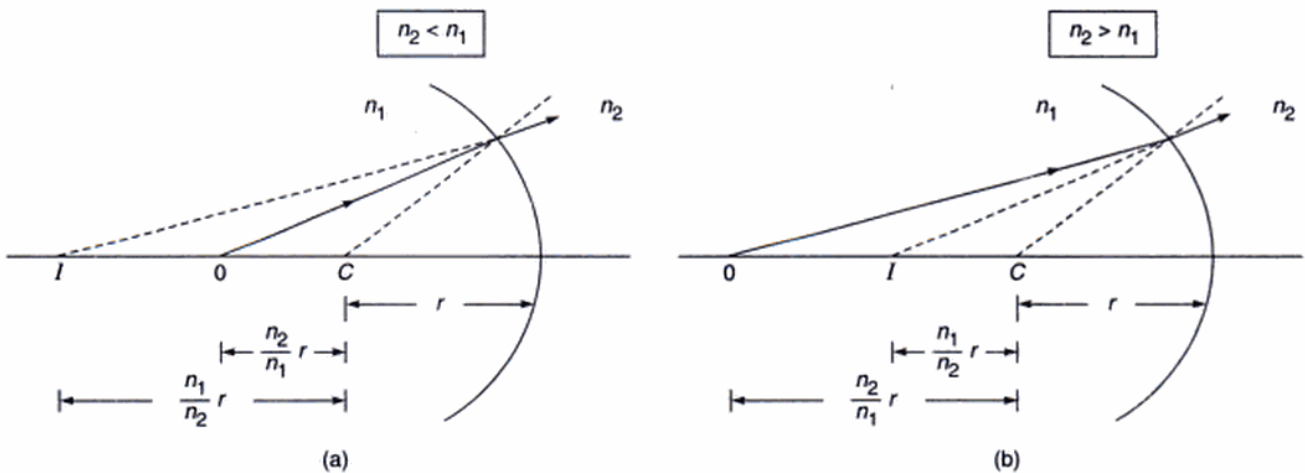


Fig. 3.12 O and I represent the aplanatic points of a spherical surface, i.e., *all* rays emanating from O appear to come from I ; (a) and (b) correspond to $n_2 < n_1$ and $n_2 > n_1$ respectively.

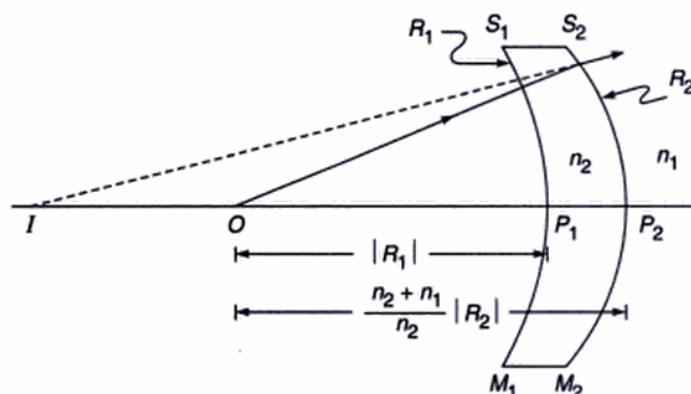


Fig. 3.13 The aplanatic lens. The object point O is at the center of curvature of the first surface $S_1P_1M_1$. The points O and I are the aplanatic points of the spherical surface $S_2P_2M_2$ —thus a perfect (virtual) image is formed at I .

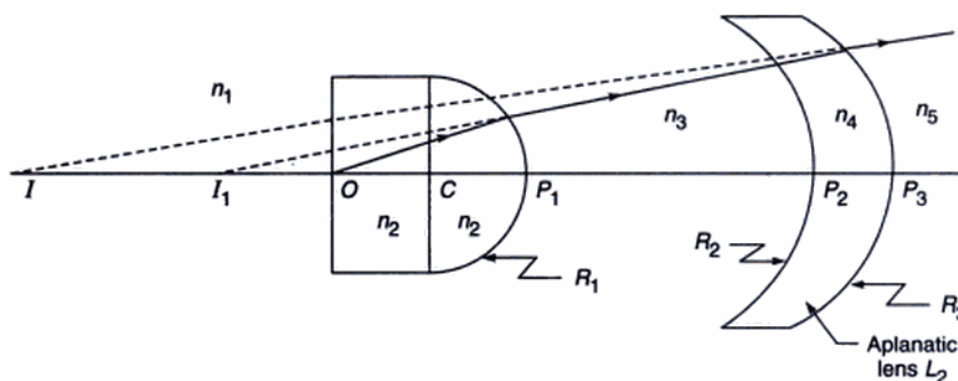


Fig. 3.14 The oil immersion objective. The points O and I_1 are the aplanatic points corresponding to the hemispherical surface of radius R_1 ; the lens L_2 acts as an aplanatic lens for the (virtual) object at I_1 .

L_2 are

$$m_1 = \frac{n_2(I_1P_1)}{n_3(OP_1)} \quad (27)$$

and

$$m_2 = \frac{n_4(IP_3)}{n_5(I_1P_3)} \quad (28)$$

Thus the oil immersion objective reduces considerably the angular divergence of the rays and results in an increase in lateral magnification without introducing spherical aberration. We should, however, mention that a perfect image is formed only of one point and therefore nearby points have some aberrations. Moreover, oil immersion objectives have a certain degree of chromatic aberration.

3.9 THE CARTESIAN OVAL

In general, for two points to form perfect images of each other, the refracting surface should not be spherical. Figure

3.15 shows the two points O and I such that all rays emanating from O (and allowed by the system) intersect each other at the other point I . Thus the curve SPM shown in Fig. 3.15 is the locus of the point S such that

$$n_1OS + n_2SI = \text{constant} \quad (29)$$

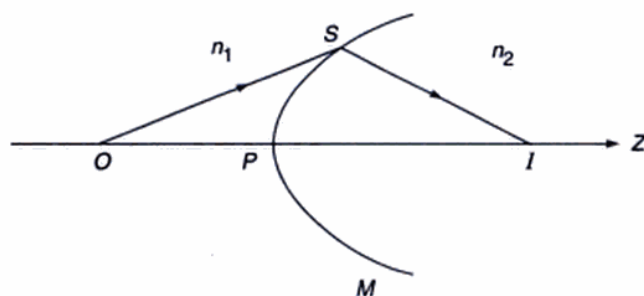


Fig. 3.15 The refracting surface (known as the Cartesian oval) is such that all rays emanating from the point O intersect at I .

The refracting surface is obtained by revolving the curve shown in the figure about the z -axis (see also Problem 2.8). The refracting surface is known as a Cartesian oval. When the object point is at infinity, the surface becomes an ellipsoid of revolution (see Problem 2.5) and under certain circumstances the surface is spherical; however, the image is then virtual [see Figs. 3.12(a) and (b)].

3.10 GEOMETRICAL PROOF FOR THE EXISTENCE OF APLANATIC POINTS

In this section we will show the existence of aplanatic points using geometrical considerations. We consider a spherical refracting surface SPM of radius r separating two media of refractive indices n_1 and n_2 (see Fig. 3.16). We will assume $n_2 < n_1$ and define

$$\mu = \frac{n_1}{n_2} \quad (30)$$

where $\mu > 1$. The point C represents the centre of the spherical surface SPM . With C as center, we draw two spheres of radii μr and r/μ as shown in Fig. 3.16. Let $IOCP$ represent any common diameter of the three spheres intersecting the outer and inner spheres at I and O respectively. From the point O , we draw an arbitrary line hitting the refracting surface at the point S . We join I and S and extend the line further as SQ . If we can show that

$$\frac{\sin \alpha}{\sin \beta} = \frac{1}{\mu} \quad (31)$$

for all values of θ_1 , then all rays emanating from the point O will appear to come from I , and O and I will be the aplanatic points for the spherical refracting surface SPM . Now,

$$\frac{IC}{CS} = \frac{\mu r}{r} = \mu \quad (32)$$

and

$$\frac{CS}{OC} = \frac{r}{r/\mu} = \mu = \frac{IC}{CS} \quad (33)$$

Thus the two triangles SOC and SIC are similar, and therefore

$$\alpha = \theta_2 \quad \text{and} \quad \beta = \angle ISC = \theta_1 \quad (34)$$

Now, considering the triangle SOC we have

$$\frac{\sin \alpha}{\sin \theta_1} = \frac{r/\mu}{r} = \frac{1}{\mu} \quad (35)$$

and using Eq. (34), we get

$$\frac{\sin \alpha}{\sin \beta} = \frac{1}{\mu} = \frac{n_2}{n_1} \quad (36)$$

proving that O and I are aplanatic points. We also have

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{\sin \beta}{\sin \alpha} = \frac{n_1}{n_2} \quad (37)$$

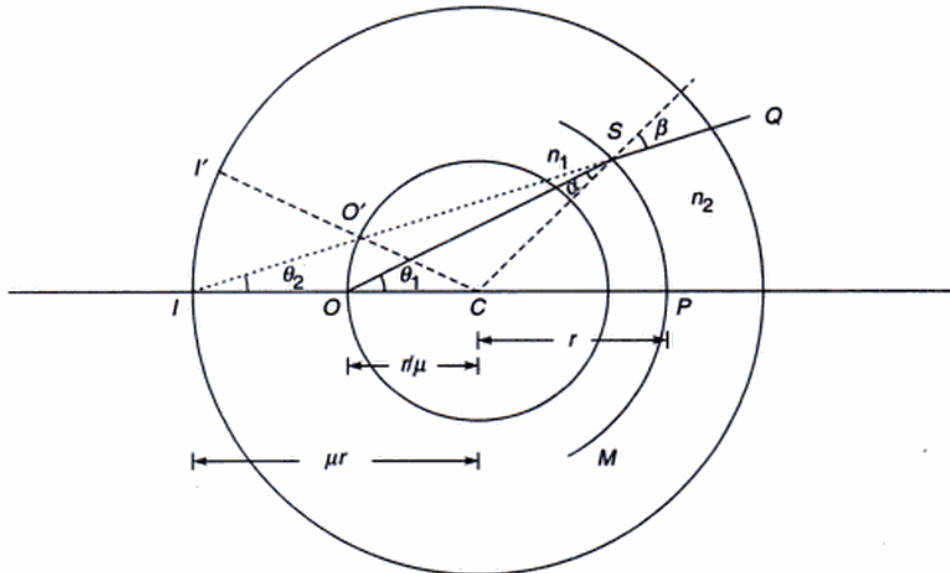


Fig. 3.16 Geometrical construction for the derivation of aplanatic points. SPM is the refracting surface of radius r . The inner and outer spheres are of radii r/μ and μr respectively. O and I are the aplanatic points.

It is obvious that the points O' and I' will also be aplanatic and therefore the image formed by a *small* planar object at O will be sharp even for the off-axis points. The system is said to be free not only from spherical aberration but also from coma. Furthermore, the linear magnification is given by

$$m = \frac{I'I}{O'O} = \frac{\mu r}{r/\mu} = \mu^2 = \left(\frac{n_1}{n_2}\right)^2 \quad (38)$$

3.11 THE SINE CONDITION

We consider a general optical system as shown in Fig. 3.17. We assume that the point O (on the axis of the system) is perfectly imaged at I , i.e. all rays emanating from O intersect each other at I . This implies that the optical system has no spherical aberration corresponding to O . We next consider a slightly off-axis point O' (directly above O) and, according to the *sine-condition*, for O' to be sharply imaged at I' we must have*

$$\frac{n_1 \sin \theta_1}{n_2 \sin \theta_2} = \frac{y_2}{y_1} = \text{linear magnification} \quad (39)$$

where θ_1 and θ_2 are defined in Fig. 3.17. Thus the linear magnification will be constant if the ratio $\sin \theta_1 / \sin \theta_2$ is constant for all points on the refracting surface and the image will be free from the aberration known as coma. It is of interest to note that according to Eq. (39) perfect imaging of (nearby) off-axis points requires a condition to be

satisfied by rays from an *on-axis* point. Also, when the condition given by Eq. (39) is satisfied, sharp imaging of a nearby point on the axis (like O_1) is *not* obtained; indeed the condition for sharp imaging of O and O_1 is quite different (see Problem 3.11).

3.11.1 Proof of the Sine Condition**

We refer to Fig. 3.17. We will assume that the axial point O is perfectly imaged at I and will use Fermat's principle to determine the condition for perfect imaging of the nearby off-axis point O' . The ray $O'B_1$ is parallel to the ray OA_1 and the ray $O'D_1$ is parallel to OC_1 . Now, since I is the image of the point O , we have

$$OPL[OA_1A_2I] = OPL[OC_1C_2I] \quad (40)$$

where *OPL* stands for the *optical path length*. Further

$$OPL[O'B_1B_2I'] = OPL[O'D_1D_2I'] \quad (41)$$

Now, the rays $O'B_1$ and OA_1 meet at infinity and therefore

$$OPL[O'B_1B_2F] = OPL[OA_1A_2F] \quad (42)$$

We next consider the triangle $FI'I'$

$$FI' = [FI^2 + |y_2|^2]^{1/2} = FI \left[1 + \frac{1}{2} \frac{|y_2|^2}{FI^2} \right]$$

Thus

$$FI' \approx FI \quad (43)$$

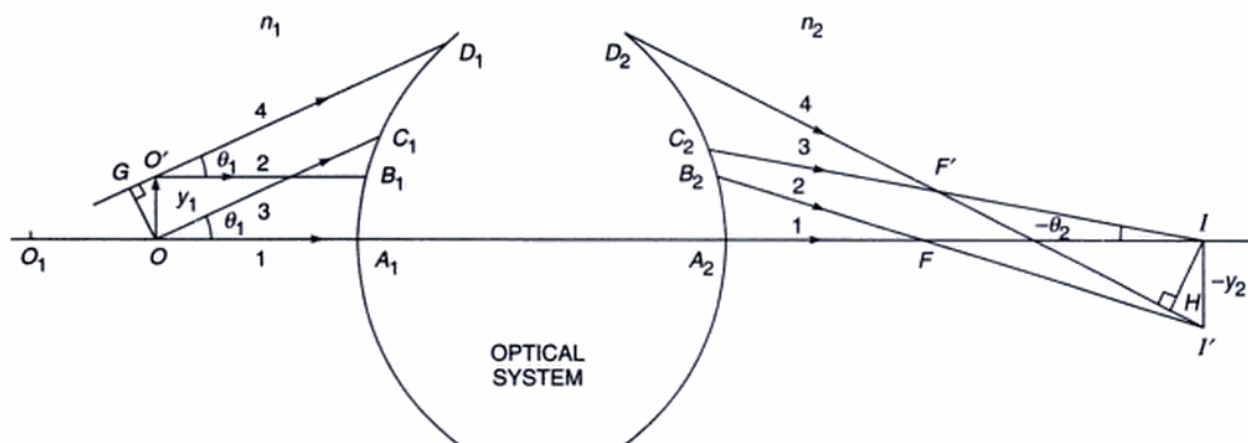


Fig. 3.17 The optical system images perfectly the points O and O' at I and I' respectively.

*It may be noted that if we use Eqs (37) and (38), we get

$$m = \frac{y_2}{y_1} = \left(\frac{n_1}{n_2}\right)^2 = \frac{n_1 \sin \theta_1}{n_2 \sin \theta_2}$$

consistent with Eq. (39).

**For a rigorous proof of the sine condition, See Ref. 3.

where we are assuming that $|y_2|$ is small enough so that terms proportional to $|y_2|^2$ can be neglected. If we add Eqs (42) and (43), we get

$$\begin{aligned} OPL[O'B_1B_2I'] &= OPL[OA_1A_2I] \\ &= OPL[OC_1C_2I] \end{aligned} \quad (44)$$

Since the left hand side of the above equation is $OPL[O'D_1D_2I']$, we get

$$OPL[O'D_1D_2I'] = OPL[OC_1C_2I] \quad (45)$$

Now, the rays 3 and 4 meet at infinity and intersect at F' , so that

$$OPL[GD_1D_2F'] = OPL[OC_1CF'] \quad (46)$$

where the point G is the foot of the perpendicular drawn from the point O on ray 4. We subtract Eq. (45) from Eq. (46) to obtain

$$OPL[F'I'] - OPL[GO'] = OPL[F'I] \quad (47)$$

or

$$n_2(F'I') - n_1(GO') = n_2(F'I)$$

or

$$n_1(GO') = n_2(F'I' - F'I) \quad (48)$$

But

$$GO' = y_1 \sin \theta_1 \quad (49)$$

and

$$F'I' - F'I \approx HI' \approx (-y_2) \sin(-\theta_2) \quad (50)$$

where H is the foot of the perpendicular from the point I on ray 4. Substituting the above two equations in Eq. (48), we get

$$\frac{n_1 \sin \theta_1}{n_2 \sin \theta_2} = \frac{y_2}{y_1} = \text{linear magnification} \quad (51)$$

showing that the linear magnification is constant if the ratio $\sin \theta_1 / \sin \theta_2$ is constant for all points on the refracting surface. The sine condition is of extensive use in the design of optical systems.

SUMMARY

- Consider refraction at a spherical surface separating two media of refractive indices n_1 and n_2 . For a point object at a distance $|u|$ on the left, the paraxial image is formed at a distance v where

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R}$$

The sign convention is as follows :

- The rays are always incident from the left on the refracting surface.
 - All distances to the right of the refracting surface are positive and distances to the left of the refracting surface are negative.
- For a thin lens of refractive index n (placed in air), let R_1 and R_2 be the radii of curvature of the left and right surfaces of the lens; then the image distance is given by

$$\frac{1}{v} - \frac{1}{u} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right)$$

which is usually referred to as the 'thin-lens formula'; the quantity f is known as the focal length of the lens.

- For a given spherical surface, there are two points for which all rays emanating from one point intersect each other at the other point. This point is at a distance equal to $n_2|R|/n_1$ from the center of the spherical surface and a virtual image is formed at a distance of $n_1|R|/n_2$ from the center. The two points are said to be the *aplanatic points* of the sphere and are utilized in the construction of aplanatic lenses.
- For two points to form perfect images of each other, the refracting surface is a Cartesian oval.

PROBLEMS

- Consider a thin biconvex lens (as shown in Fig. 3.18) made of a material whose refractive index is 1.5. The radii of curvature of the first and second surfaces (R_1 and R_2) are +100 and -60 cm respectively. The lens is placed in air (i.e. $n_1 = n_3 = 1$). For an object at a distance of 100 cm from the lens, determine the position and linear magnification of the (paraxial) image. Also calculate x_1 and x_2 and verify Newton's formula [Eq. (20)].
[Ans: $x_1 = -25$ cm and $x_2 = +225$ cm]
 - Repeat the calculations of the above problem when the object is at a distance of 50 cm.
- Consider a thin lens (made of a material of refractive index n_2) having different media on the two sides; let n_1 and n_3 be the refractive indices of the media on the left and on the right of the lens respectively. Using Eq. (5) and considering successive refractions at the two surfaces, derive Eq. (14).
- Referring again to Fig. 3.18 assume a biconvex lens with $|R_1| = 100$ cm, $|R_2| = 60$ cm with $n_1 = 1.0$ but $n_3 = 1.6$. For $u = -50$ cm determine the position of the

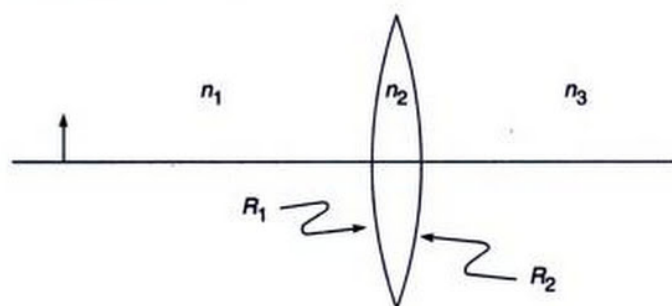


Fig. 3.18

(paraxial) image. Also determine the first and second principal foci and verify Newton's formula. Draw the ray diagram.

[Ans: $x_1 = 250$ cm, $x_2 = 576$ cm]

- 3.4 (a) In Fig. 3.18, assume the convex lens to be replaced by a (thin) biconcave lens with $|R_1| = 100$ cm and $|R_2| = 60$ cm. Assume $n_1 = n_3 = 1$ and $n_2 = 1.5$. Determine the position of the image and draw an approximate ray diagram for $u = -100$ cm.

- (b) In (a), assume $n_1 = n_3 = 1.5$ and $n_2 = 1.3$. Repeat the calculations and draw the ray diagram. What is the qualitative difference between the systems in (a) and (b)?

- 3.5 Consider an object of height 1 cm placed at a distance of 24 cm from a convex lens of focal length 15 cm (see Fig. 3.19). A concave lens of focal length -20 cm is placed beyond the convex lens at a distance of 25 cm. Draw the ray diagram and determine the position and size of the final image.

[Ans: Real image at a distance of 60 cm from the concave lens.]

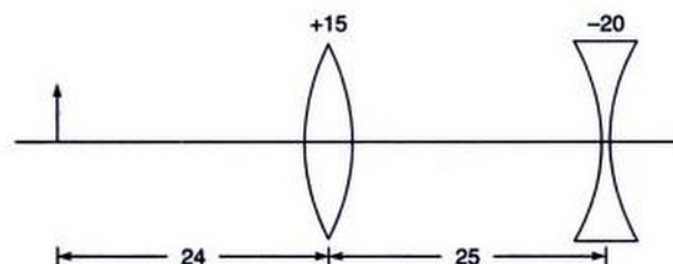


Fig. 3.19 An optical system consisting of a thin convex and a thin concave lens. All distances are measured in centimeters.

- 3.6 Consider a thick biconvex lens whose magnitude of the radii of curvature of the first and second surfaces are 45 and 30 cm respectively. The thickness of the lens is 5 cm and the refractive index of the material, it is made of, is 1.5. For an object of height 1 cm at

distance of 90 cm from the first surface, determine the position and size of the image. Draw the ray diagram for the axial point of the object.

[Ans: Real image at a distance of 60 cm from the second surface.]

- 3.7 In Problem 3.6 assume that the second surface is silvered so that it acts like a concave mirror. For an object of height 1 cm at a distance of 90 cm from the first surface determine the position and size of the image and draw the ray diagram.

[Ans: Real image at a distance of about 6.2 cm from the first surface. (Remember the sign convention.)]

- 3.8 Consider a sphere of radius 20 cm of refractive index 1.6 (see Fig. 3.20). Show that the paraxial focal point is at a distance of 6.7 cm from the point P_2 .

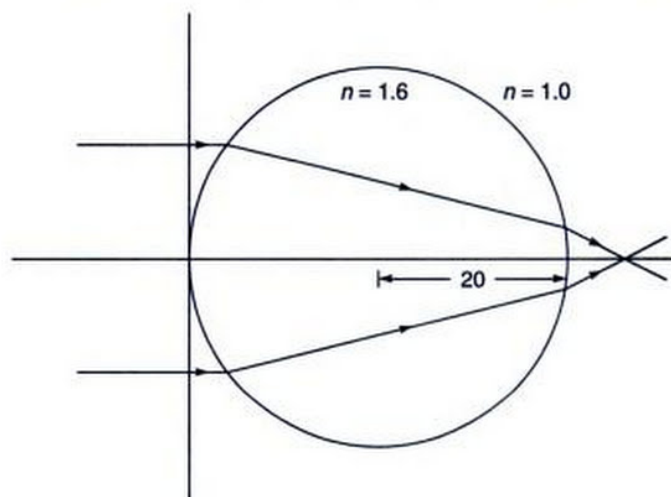


Fig. 3.20

- 3.9 Consider a hemisphere of radius 20 cm and refractive index 1.5. Show that parallel rays will focus at a point 40 cm from P_2 (see Fig. 3.21).

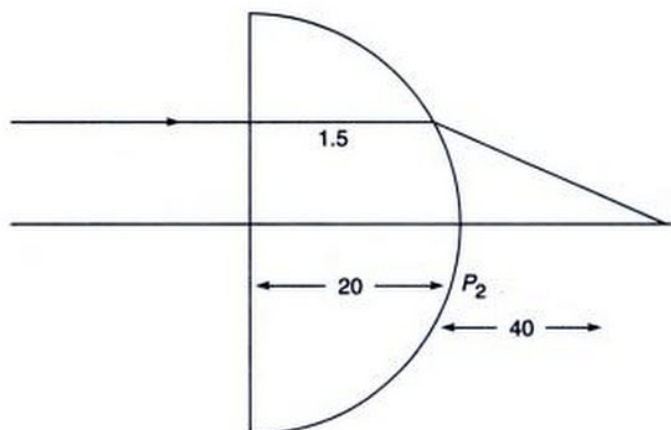


Fig. 3.21

- 3.10** Consider a lens of thickness 1 cm, made of a material of refractive index 1.5, placed in air. The radii of curvature of the first and second surfaces are +4 cm and -4 cm respectively. Determine the point at which parallel rays will focus.

[Ans: At a distance of about 4.55 cm from the second surface.]

- 3.11** Show that, for two points O and O_1 (on the axis of the optical system) to be imaged perfectly at I and I_1 (see Fig. 3.22), the following condition has to be satisfied

$$\frac{n_1 \sin^2 \frac{1}{2} \theta_1}{n_2 \sin^2 \frac{1}{2} \theta_2} = \frac{I_1 I}{O_1 O}$$

which is known as Herschel's condition.

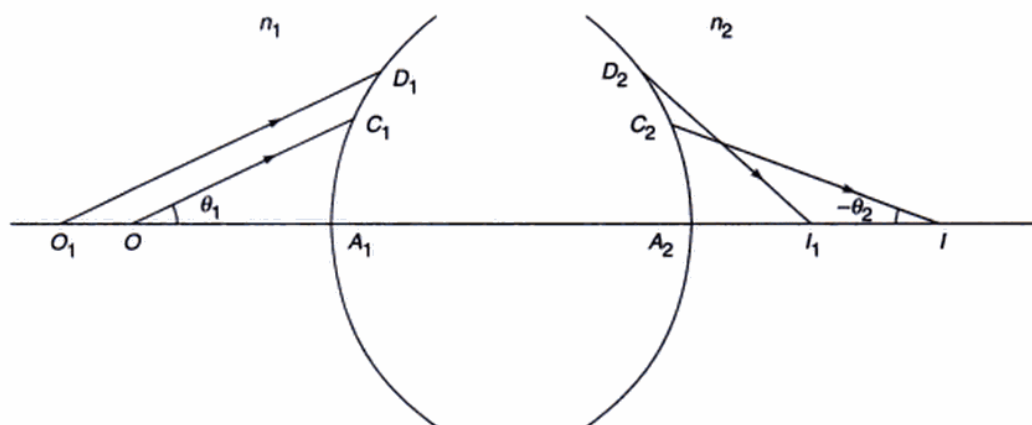


Fig. 3.22 The optical system images perfectly the points O and O_1 at I and I_1 respectively.

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
2. R.P. Feynman, R.B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass, 1965.
3. A.K. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. [Reprinted by Macmillan India, New Delhi.]
4. R.H. Penfield, 'Consequences of parameter invariance in geometrical optics', *American Journal of Physics*, Vol. 24, 19, 1956.

Chapter 4

The Matrix Method in Paraxial Optics

In dealing with a system of lenses we simply chase the ray through the succession of lenses. That is all there is to it.

—Richard Feynman in
Feynman Lectures in Physics

4.1 INTRODUCTION*

Let us consider a ray PQ incident on a refracting surface SQS' separating two media of refractive indices n_1 and n_2 (see Fig. 4.1). Let NQN' denote the normal to the surface. The direction of the refracted ray is completely determined from the following conditions:

- the incident ray, the refracted ray and the normal lie in the same plane; and
- if θ_1 and θ_2 represent the angles of incidence and refraction respectively, then

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{n_2}{n_1} \quad (1)$$

Optical systems, in general, are made up of a large number of refracting surfaces (like in a combination of lenses) and

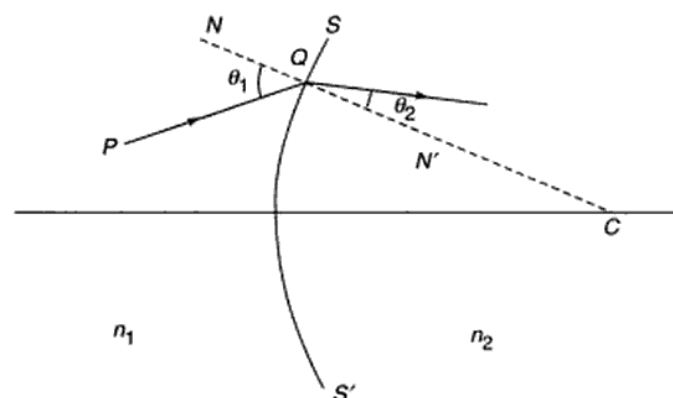


Fig. 4.1 Refraction of a ray by a surface SQS' which separates two media of refractive indices n_1 and n_2 ; NQN' denotes the normal at the point Q . If the refracting surface is spherical then the normal NQN' will pass through the centre of curvature C .

any ray can be traced through the system by using the above conditions. In order to obtain the position of the final image due to such a system, one has to calculate step-by-step the position of the image due to each surface and this image will act as an object for the next surface. Such a step-by-step analysis becomes complicated as the number of elements of an optical system increases. We shall, in this chapter, develop the matrix method which can be applied with ease under such situations. This method indeed lends itself to direct use in the computers for tracing rays through complicated optical systems.

Before we describe the matrix formulation of geometric optics it is necessary to mention the rule of matrix multiplication and the use of matrices for solving linear equations. A $(m \times n)$ matrix has m rows and n columns and has $(m \times n)$ elements; thus the matrix

$$A = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \quad (2)$$

has 2 rows and 3 columns and has $2 \times 3 = 6$ elements. A $(m \times n)$ matrix can be multiplied only to a $(n \times p)$ matrix to obtain a $m \times p$ matrix. Let

$$B = \begin{pmatrix} g \\ h \\ i \end{pmatrix} \quad (3)$$

represent a (3×1) matrix. Then the product

$$AB = \begin{pmatrix} a & b & c \\ d & e & f \end{pmatrix} \begin{pmatrix} g \\ h \\ i \end{pmatrix} = \begin{pmatrix} (ag + bh + ci) \\ (dg + eh + fi) \end{pmatrix} \quad (4)$$

will be a (2×1) matrix, and the product BA has no meaning.

If we define a (2×3) matrix

$$A' = \begin{pmatrix} a' & b' & c' \\ d' & e' & f' \end{pmatrix}$$

* The author thanks Professor K. Thyagarajan for his help in writing this chapter.

then $A' = A$

if and only if $a' = a$, $b' = b$, $c' = c$, $d' = d$, $e' = e$ and $f' = f$, i.e. all the elements must be equal. The set of two equations

$$\begin{cases} x_1 = ay_1 + by_2 \\ x_2 = cy_1 + dy_2 \end{cases} \quad (5)$$

can be written in the following form:

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} ay_1 + by_2 \\ cy_1 + dy_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad (6)$$

the last step follows from the rule of matrix multiplication.

Further, if we have

$$\text{and} \quad \begin{cases} y_1 = ez_1 + fz_2 \\ y_2 = gz_1 + hz_2 \end{cases} \quad (7)$$

then

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad (8)$$

Consequently,

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \quad (9)$$

$$\text{or} \quad X = BZ, \quad (10)$$

where X and Z represent (2×1) matrices:

$$X \equiv \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \quad Z \equiv \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}, \quad (11)$$

and B represents a (2×2) square matrix

$$\begin{aligned} B &= \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} \\ &= \begin{bmatrix} [ae + bg] & [af + bh] \\ [ce + dg] & [cf + dh] \end{bmatrix} \end{aligned} \quad (12)$$

Equations (9) and (12) tell us that

$$\text{and} \quad \begin{cases} x_1 = (ae + bg)z_1 + (af + bh)z_2 \\ x_2 = (ce + dg)z_1 + (cf + dh)z_2 \end{cases} \quad (13)$$

which can be verified by direct substitution. We will now use the matrix method to trace paraxial rays through a cylindrically symmetric optical system.

4.2 THE MATRIX METHOD

We will consider a cylindrically symmetric optical system similar to the one shown in Fig. 4.2. The axis of symmetry is chosen as the z -axis. We will be considering only paraxial rays in this chapter; non-paraxial rays lead to what are known as aberrations which will be discussed in Chapter 5.

In the paraxial approximation we may confine ourselves to rays which pass through the axis of the system; these rays remain confined to a single plane. Such a ray can be specified by its distance from the axis of the system and the angle made by the ray with the axis; for example, in Fig. 4.2, the point P on the ray is at a distance x_1 from the axis and makes an angle α_1 with the axis. The quantities (x_1, α_1) represent the coordinates of the ray. However, instead of specifying the angle made by the ray with the z -axis, we will specify the quantity.

$$\lambda = n \cos \psi (= n \sin \alpha)$$

which represents the product of the refractive index with the sine of the angle that the ray makes with the z -axis this quantity is known as the optical direction cosine.

Now, when a ray propagates through an optical system, it undergoes only two operations: (a) translation and (b) refraction. The rays undergo translation when they propagate through a homogeneous medium as in the region PQ (see Fig. 4.2). However, when it strikes an interface of two media, it undergoes refraction. We will now study the

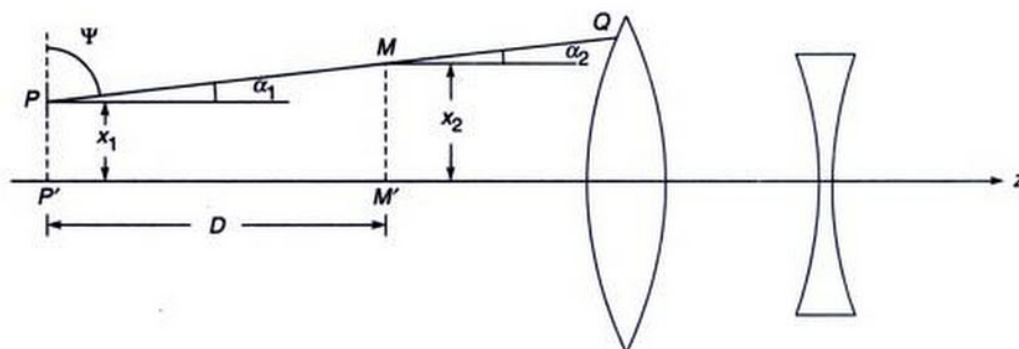


Fig. 4.2 In a homogeneous medium the ray travels in a straight line.

effect of translation and of refraction on the coordinates of the ray.

(a) Effect of Translation

Consider a ray travelling in a homogeneous medium of refractive index n_1 which is initially at a distance x_1 from the z -axis and makes an angle α_1 with the axis (see point P in Fig. 4.2). Let (x_2, α_2) represent the coordinates of the ray at the point M (see Fig. 4.2). Since the medium is homogeneous, the ray travels in a straight line and, therefore,

$$\alpha_2 = \alpha_1 \quad (14)$$

Further, if PP' and MM' are perpendiculars on the axis and if $P'M' = D$, then

$$x_2 = x_1 + D \tan \alpha_1 \quad (15)$$

Since we are interested only in paraxial rays, α_1 is very small and hence we can make use of the approximation $\tan \alpha_1 \approx \alpha_1$, where α_1 is measured in radians. Thus, Eq. (15) reduces to

$$x_2 = x_1 + \alpha_1 D \quad (16)$$

$$\text{If } \lambda_1 = n_1 \alpha_1 \quad (17)$$

$$\text{and } \lambda_2 = n_2 \alpha_2 \quad (18)$$

then, using Eqs (15) and (17), we get

$$\text{and } \left. \begin{aligned} \lambda_2 &= \lambda_1 \\ x_2 &= x_1 + \frac{D}{n_1} \lambda_1 \end{aligned} \right\} \quad (19)$$

which may be combined into the following matrix equation:

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ D/n_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (20)$$

Thus, if a ray is initially specified by a (2×1) matrix with elements λ_1 and x_1 , then the effect of translation through a distance D in a homogeneous medium of refractive index n_1 , is completely given by the 2×2 matrix

$$T = \begin{pmatrix} 1 & 0 \\ D/n_1 & 1 \end{pmatrix} \quad (21)$$

and the final ray is given by Eq. (20). The matrix T is known as the translation matrix. Notice that

$$\det T = \begin{vmatrix} 1 & 0 \\ D/n_1 & 1 \end{vmatrix} = 1 \quad (22)$$

(b) Effect of Refraction

We will now determine the matrix which would represent the effect of refraction through a spherical surface of radius

of curvature R . Consider the ray AP intersecting a spherical surface (separating two media of refractive indices n_1 and n_2 respectively) at the point P and getting refracted along PB (see Fig. 4.3). If θ_1 and θ_2 are the angles made by the incident and the refracted ray with the normal to the surface at P (i.e., with the line joining P to the centre of curvature C), then according to Snell's law

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \quad (23)$$

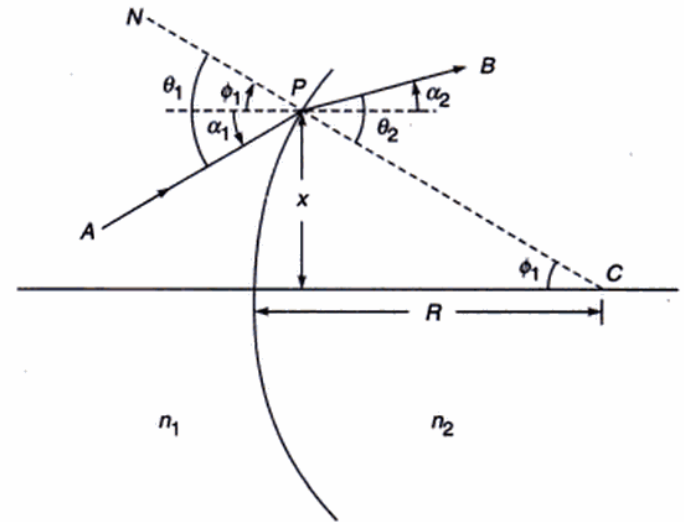


Fig. 4.3 The refraction of a ray at a spherical surface.

Since we are dealing with paraxial rays, one can make use of the approximation $\sin \theta \approx \theta$. Thus Eq. (23) reduces to

$$n_1 \theta_1 \approx n_2 \theta_2 \quad (24)$$

From Fig. 4.3 it follows that

$$\theta_1 = \phi_1 + \alpha_1 \quad \text{and} \quad \theta_2 = \phi_1 + \alpha_2 \quad (25)$$

where α_1 , α_2 and ϕ_1 are respectively the angles that the incident ray, the refracted ray and the normal to the surface make with the z -axis. Also, since ϕ_1 is small, we may write

$$\phi_1 = \frac{x}{R} \quad (26)$$

Now, from Eqs (24) and (25), we get

$$n_1(\phi_1 + \alpha_1) \approx n_2(\phi_1 + \alpha_2)$$

or

$$n_2 \alpha_2 \approx n_1 \alpha_1 - \frac{n_2 - n_1}{R} x \quad (27)$$

where we have used Eq. (26). Thus

$$\lambda_2 = \lambda_1 - Px \quad (28)$$

where

$$P = \frac{n_2 - n_1}{R} \quad (29)$$

is known as the power of the refracting surface. Also, since the height of the ray at P , before and after refraction, is the same (i.e. $x_2 = x_1$) we obtain, for the refracted ray,

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (30)$$

Thus, refraction through a spherical surface can be characterized by a 2×2 matrix:

$$\mathcal{R} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \quad (31)$$

It may be noted here that

$$\det \mathcal{R} = \begin{vmatrix} 1 & -P \\ 0 & 1 \end{vmatrix} = 1 \quad (32)$$

In general, an optical system made up of a series of lenses, can be characterized by the refraction and translation matrices.

If a ray is specified by $\begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$ when it enters an optical system, and is specified by $\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix}$ when it leaves the system, then one can, in general, write

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (33)$$

where the matrix

$$S = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \quad (34)$$

is called the system matrix and is determined solely by the optical system. The negative signs in some of the elements of S have been chosen for convenience. Since the only two operations a ray undergoes in traversing through an optical system are refraction and translation, the system matrix is, in general, a product of refraction and translation matrices. Also, using the property that the determinant of the product of matrices is the product of the determinant of the matrices, we obtain

$$\det S = 1 \quad (35)$$

i.e.,

$$bc - ad = 1 \quad (36)$$

We should mention here that the quantities b and c are dimensionless. The quantities a and P have the dimension of inverse length and the quantity d has the dimension of length. In general, the units will not be given; however, it will be implied that a and P are in cm^{-1} and d is in cm .

4.2.1 Imaging by a Spherical Refracting Surface

As a simple illustration of the use of the matrix method we consider imaging by a spherical surface separating two media of refractive indices n_1 and n_2 (see Fig. 4.4); the same problem was discussed in the previous chapter using the standard geometrical method. Let (λ_1, x_1) , (λ', x') , (λ'', x'') and (λ_2, x_2) represent the coordinates of the ray at O , A' (just before refraction), A'' (just after refraction) and at I respectively.

We will be using the analytical geometry sign convention so that the co-ordinates on the left of the point P are negative and co-ordinates on the right of P are positive (see Sec. 3.2.1). Thus

$$\begin{pmatrix} \lambda' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -u/n_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

$$\begin{pmatrix} \lambda'' \\ x'' \end{pmatrix} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda' \\ x' \end{pmatrix}$$

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ v/n_2 & 1 \end{pmatrix} \begin{pmatrix} \lambda'' \\ x'' \end{pmatrix}$$

or

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ v/n_2 & 1 \end{pmatrix} \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -u/n_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

Simple manipulations give

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 + \frac{Pu}{n_1} & -P \\ \frac{v}{n_2} \left(1 + \frac{Pu}{n_1} \right) - \frac{u}{n_1} & \left(1 - \frac{vP}{n_2} \right) \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (37)$$

from which we obtain

$$x_2 = \left[\frac{v}{n_2} \left(1 + \frac{Pu}{n_1} \right) - \frac{u}{n_1} \right] \lambda_1 + \left[1 - \frac{vP}{n_2} \right] x_1 \quad (38)$$

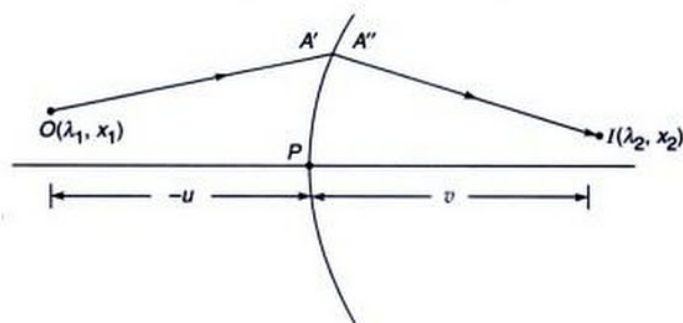


Fig. 4.4 Imaging by a spherical refracting surface separating two media of refractive indices n_1 and n_2 .

For a ray emanating from an axial object point (i.e. for $x_1 = 0$) the image plane is determined by the condition $x_2 = 0$. Thus in the above equation, the coefficient of λ_1 should vanish and therefore

$$\frac{u}{n_1} = \frac{v}{n_2} \left(1 + \frac{Pu}{n_1} \right)$$

or

$$\frac{n_2}{v} - \frac{n_1}{u} = P = \frac{n_2 - n_1}{R} \quad (39)$$

which is the same as derived in the previous chapter. Hence, on the image plane

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 + \frac{Pu}{n_1} & -P \\ 0 & 1 - \frac{vP}{n_2} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (40)$$

giving

$$x_2 = \left(1 - \frac{vP}{n_2} \right) x_1$$

Thus the magnification is given by

$$m = \frac{x_2}{x_1} = 1 - \frac{vP}{n_2}$$

which on using Eq. (39) gives

$$m = \frac{n_1 v}{n_2 u}$$

consistent with Eq. (23) of the previous chapter.

4.2.2 Imaging by a Co-axial Optical System

We will next derive the position of the image plane for an object plane, which is at distance $-D_1$ from the first refracting surface of the optical system (see Fig. 4.5). Let the image be formed at a distance D_2 from the last refracting surface. Now, according to our sign convention, for points on the left of a refracting surface, the distances will be negative and for points on the right of the refracting surface the distances will be positive, thus D_1 is an intrinsically negative quantity. Further, if D_2 is found to be positive the image is real and is formed on the right of the refracting surface; on the other hand, if D_2 is found to be negative, the image will be virtual and will be formed on the left of the last refracting surface.

Let us consider a ray $O'P$ starting from the point O' which lies in the object plane. Let QI' be the ray emerging from the last surface; the point I' is assumed to lie on the image plane—see Fig. 4.5 (the point I is the paraxial image of the point O and the image plane is defined to be the plane which contains the point I and is normal to the axis). Let (λ_1, x_1) , (λ', x') , (λ'', x'') and (λ_2, x_2) represent the coordinates of the ray at O' , P , Q and I' respectively. Then

$$\begin{cases} \begin{pmatrix} \lambda' \\ x' \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ -D_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \\ \begin{pmatrix} \lambda'' \\ x'' \end{pmatrix} = \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} \lambda' \\ x' \end{pmatrix} \\ \begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ D_2 & 1 \end{pmatrix} \begin{pmatrix} \lambda'' \\ x'' \end{pmatrix} \end{cases}$$

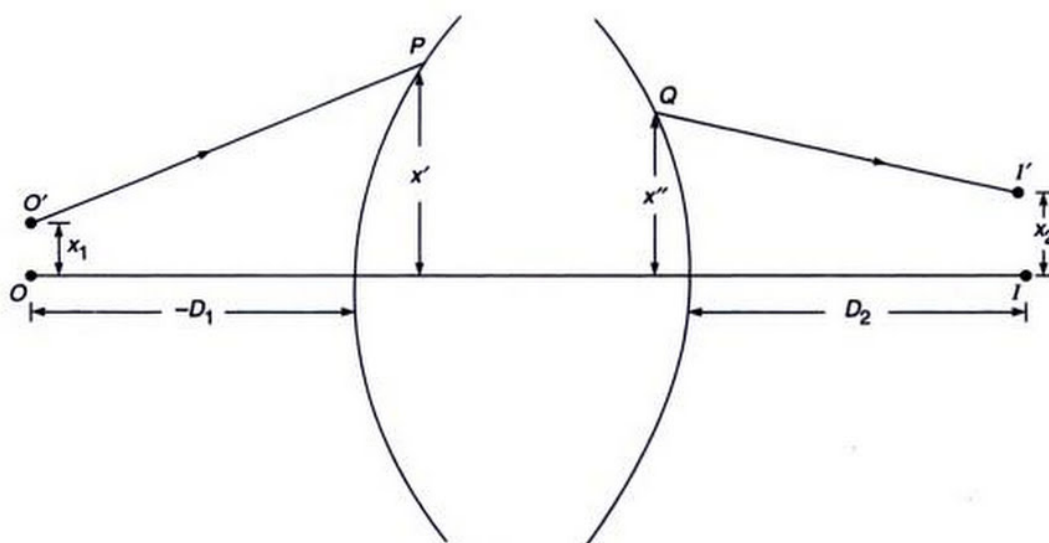


Fig. 4.5 The object point O is at a distance $(-D_1)$ from the first refracting surface. The paraxial image is assumed to be formed at a distance D_2 from the last refracting surface.

Thus

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ D_2 & 1 \end{pmatrix} \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} \begin{pmatrix} 1 & 0 \\ -D_1 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (41)$$

where the first and the third matrices on the RHS correspond to translations by distances D_2 and $(-D_1)$ respectively (in a medium of refractive index unity); the second matrix correspond to the system matrix of the optical system. Carrying out the matrix multiplications, we obtain

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} b + aD_1 & -a \\ bD_2 + aD_1D_2 - cD_1 - d & c - aD_2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (42)$$

Thus

$$x_2 = (bD_2 + aD_1D_2 - cD_1 - d)\lambda_1 + (c - aD_2)x_1$$

For a ray emanating from the axial object point (i.e. for $x_1 = 0$) the image plane is determined by the condition $x_2 = 0$. Thus, for the image plane we must have

$$bD_2 + aD_1D_2 - cD_1 - d = 0 \quad (43)$$

which would give us the relationship between the distances D_1 and D_2 . Thus, corresponding to the image plane, we have

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} b + aD_1 & -a \\ 0 & c - aD_2 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (44)$$

For $x_2 \neq 0$, we obtain

$$x_2 = (c - aD_2)x_1$$

Consequently, the magnification of the system, $M \left(= \frac{x_2}{x_1} \right)$

would be given by

$$M = \frac{x_2}{x_1} = c - aD_2 \quad (45)$$

Further, since

$$\begin{vmatrix} b + aD_1 & -a \\ 0 & c - aD_2 \end{vmatrix} = 1$$

we obtain

$$b + aD_1 = \frac{1}{c - aD_2} = \frac{1}{M} \quad (46)$$

Hence, if x_1 and x_2 correspond to object and image planes, then for a general optical system we may write

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1/M & -a \\ 0 & M \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (47)$$

Example 4.1 Obtain the system matrix for a thick lens and derive the thin lens and thick lens formulae.

Solution: Let us consider a lens of thickness t and made of a material of relative refractive index n (see Fig. 4.6). Let R_1 and R_2 be the radii of curvatures of the two surfaces. The ray is assumed to strike the first surface of the lens at P and emerge from the point Q ; let the coordinates of the ray at P and Q be

$$\begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} \quad (48)$$

where λ_1 and λ_2 are the optical direction cosines of the ray at P and Q , x_1 and x_2 are the distances of the points P and Q from the axis (see Fig. 4.6). The ray, in propagating from P to Q , undergoes two refractions [one at the first surface (whose radius of curvature is R_1) and the other at the second surface (whose radius of curvature is R_2)] and a translation through a distance* t in a medium of refractive index n . Thus

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 & -P_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t/n & 1 \end{pmatrix} \begin{pmatrix} 1 & -P_1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix} \quad (49)$$

where

$$P_1 = \frac{n-1}{R_1} \quad \text{and} \quad P_2 = \frac{1-n}{R_2} = -\frac{n-1}{R_2} \quad (50)$$

represent the powers of the two refracting surfaces. Thus our system matrix is given by

$$\begin{aligned} S &= \begin{pmatrix} b & -a \\ -d & c \end{pmatrix} = \begin{pmatrix} 1 & -P_2 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t/n & 1 \end{pmatrix} \begin{pmatrix} 1 & -P_1 \\ 0 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 1 - \frac{P_2 t}{n} & -P_1 - P_2 \left(1 - \frac{t}{n} P_1\right) \\ \frac{t}{n} & 1 - \frac{t}{n} P_1 \end{pmatrix} \quad (51) \end{aligned}$$

For a thin lens, $t \rightarrow 0$ and the system matrix takes the following form:

$$S = \begin{pmatrix} 1 & -P_1 - P_2 \\ 0 & 1 \end{pmatrix} \quad (52)$$

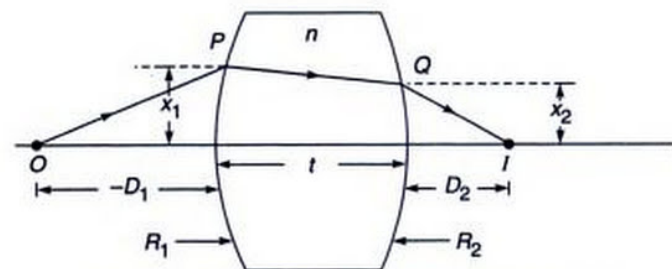


Fig. 4.6 A paraxial ray passing through a thick lens of thickness t .

* Notice that since we are dealing with paraxial rays, the distance between P and Q is approximately t .

Thus for a thin lens,

$$a = P_1 + P_2, b = 1, c = 1 \text{ and } d = 0 \quad (53)$$

Substituting the above values of a, b, c and d in Eq. (43), we obtain

$$D_2 + (P_1 + P_2) D_1 D_2 - D_1 = 0,$$

or

$$\begin{aligned} \frac{1}{D_2} - \frac{1}{D_1} &= (P_1 + P_2) \\ &= (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \end{aligned} \quad (54)$$

or

$$\frac{1}{D_2} - \frac{1}{D_1} = \frac{1}{f} \quad (55)$$

where

$$f = \frac{1}{P_1 + P_2} = \left[(n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \right]^{-1} \quad (56)$$

represents the focal length of the lens. Equation (55) is the well-known thin lens formula. (The signs of R_1 and R_2 for different kinds of lenses are shown in Fig. 4.7). Thus the system matrix for a thin lens is given by

$$S = \begin{pmatrix} 1 & -\frac{1}{f} \\ 0 & 1 \end{pmatrix} \quad (57)$$

For a thick lens, we have from Eq. (51)

$$\left. \begin{aligned} a &= P_1 + P_2 \left(1 - \frac{t}{n} P_1 \right), \quad b = 1 - \frac{P_2 t}{n} \\ c &= 1 - \frac{t}{n} P_1, \quad d = -\frac{t}{n} \end{aligned} \right\} \quad (58)$$

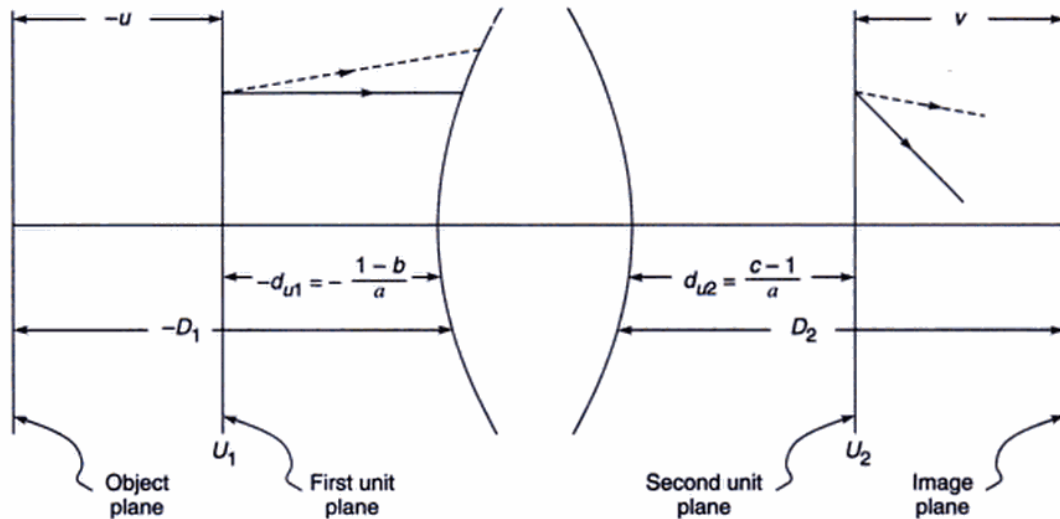


Fig. 4.8 U_1 and U_2 are the two unit planes. A ray emanating at any height from the first unit plane will cross the second unit plane at the same height.

*Obviously, if we consider U_1 as an object plane, then U_2 is the corresponding image plane.

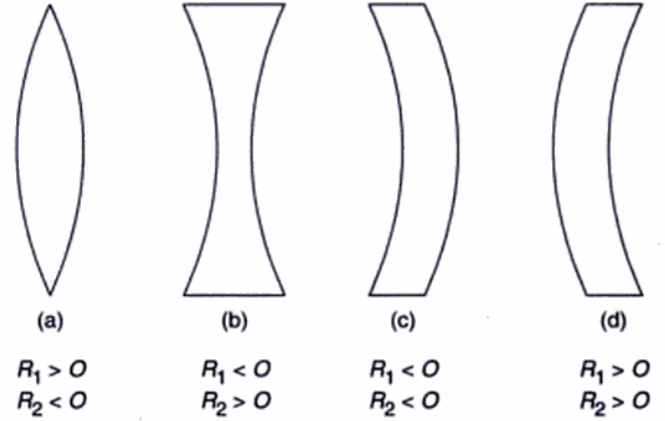


Fig. 4.7 Signs of R_1 and R_2 for different lens types.

If we substitute the above values for a, b, c and d in Eq. (43), we get the required relation between D_1 and D_2 ; however, for thick lenses it is more convenient to define the unit and the nodal planes which we shall do in the following sections.

4.3 UNIT PLANES

The unit planes are two planes, one each in the object and the image space, between which the magnification M is unity, i.e., any paraxial ray emanating from the unit plane in the object space will emerge at the same height from the unit plane in the image space. Thus, if d_{u1} and d_{u2} represent the distances of the unit planes from the refracting surfaces (see Fig. 4.8)* we obtain from Eq. (46):

$$b + ad_{u1} = \frac{1}{c - ad_{u2}} = 1 \quad (59)$$

$$\text{or} \quad d_{u1} = \frac{1-b}{a} \quad (60)$$

$$d_{u2} = \frac{c-1}{a} \quad (61)$$

Hence the unit planes are determined completely by the elements of the system matrix S .

It will be convenient to measure distances from the unit planes. Thus if u is the distance of the object plane from the first unit plane and v is the distance of the corresponding image plane from the second unit plane (see Fig. 4.8), we would obtain

$$D_1 = u + d_{u1} = u + \frac{1-b}{a} \quad (62)$$

and

$$D_2 = v + d_{u2} = v + \frac{c-1}{a} \quad (63)$$

Now, from Eq. (43) we have

$$D_2 = \frac{d + cD_1}{b + aD_1} \quad (64)$$

Substituting for D_1 and D_2 from Eqs (62) and (63), we get

$$\begin{aligned} v + \frac{c-1}{a} &= \frac{d + cu + \frac{c(1-b)}{a}}{b + au + \frac{(1-b)}{a}} \\ \text{or} \quad v &= \frac{ad - bc + c(au + 1) - (c-1)(1+au)}{a(1+au)} \\ &= \frac{au}{a(1+au)} \end{aligned} \quad (65)$$

where we have used the condition that

$$\det S = bc - ad = 1 \quad (66)$$

On simplification, we obtain

$$\frac{1}{v} - \frac{1}{u} = a \quad (67)$$

Thus $1/a$ represents the focal length of the system if the distances are measured from the two unit planes. For example, for a thick lens one obtains [using Eqs (58), (60) and (61)]:

$$d_{u1} = \frac{P_2 t}{n} \left[\frac{1}{P_1 + P_2 \left(1 - \frac{t}{n} P_1 \right)} \right] \quad (68)$$

and

$$d_{u2} = -\frac{t}{n} \left[\frac{P_1}{P_1 + P_2 \left(1 - \frac{t}{n} P_1 \right)} \right] \quad (69)$$

For a thick double convex lens with $|R_1| = |R_2|$

$$P_1 = P_2 = \frac{n-1}{R} \quad (70)$$

where $R = |R_1| = |R_2|$. Thus

$$d_{u1} = \frac{t}{n} \left[\frac{1}{2 - \frac{t}{n} \frac{n-1}{R}} \right] \approx \frac{t}{2n} \quad (71)$$

and

$$d_{u2} = -\frac{t}{n} \left[\frac{1}{2 - \frac{t}{n} \frac{n-1}{R}} \right] \approx -\frac{t}{2n} \quad (72)$$

where we have assumed $t \ll R$ which is indeed the case for most thick lenses. The positions of the unit planes are shown in Fig. 4.9. In order to calculate the focal length we note from Eq. (67) that

$$\frac{1}{f} = a = P_1 + P_2 \left(1 - \frac{t}{n} P_1 \right) \quad (73)$$

where we have used Eq. (58). Thus

$$\frac{1}{f} = (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) + \frac{(n-1)^2 t}{n R_1 R_2} \quad (74)$$

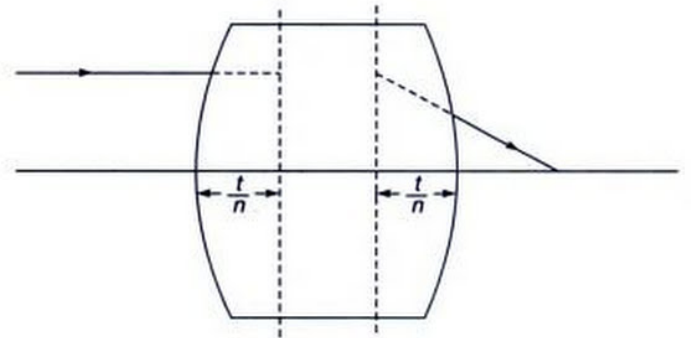


Fig. 4.9 Unit planes of a thick biconvex lens.

4.4 NODAL PLANES

Nodal points are two points on the axis which have a relative angular magnification of unity, i.e. a ray striking the first point at an angle α emerges from the second point at the same angle (see Fig. 4.10). The planes which pass through these points and are normal to the axis are known as nodal planes.

To determine the position of the nodal points, we consider two axial points N_1 and N_2 at distances d_{n1} and d_{n2} from the two refracting surfaces respectively (see Fig. 4.10). From the definition of nodal points, we require that a ray incident at an angle α_1 on the point N_1 emerge from the optical system at the same angle α_1 from the other point N_2 .

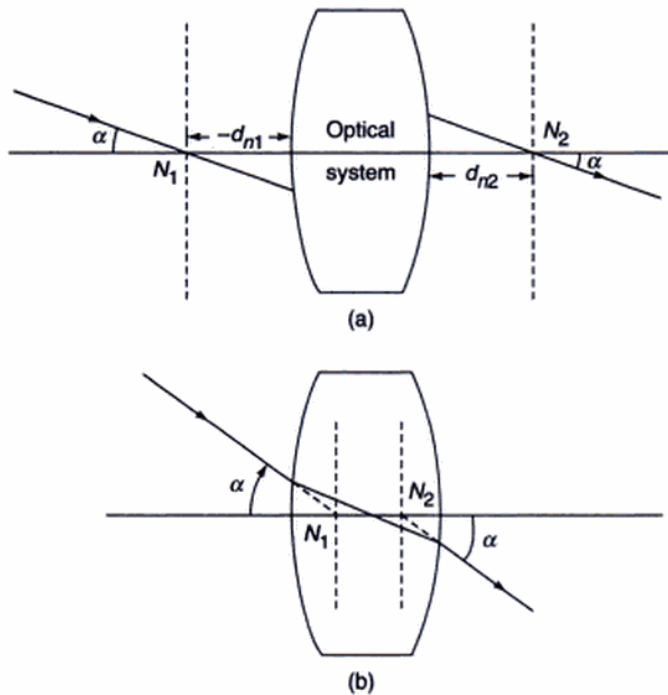


Fig. 4.10 N_1 and N_2 denote the two nodal points of an optical system. The nodal points can also lie inside the optical system as shown in (b).

Since we have assumed the media on either side of the system to have the same refractive index, this condition requires the equality of λ_1 and λ_2 . Also, since we are considering an axial object point, $x_1 = 0$, we get from Eq. (44)

$$\lambda_2 = (b + ad_{n1})\lambda_1 = \lambda_1 \quad (75)$$

Thus

$$b + ad_{n1} = 1 \quad (76)$$

or

$$d_{n1} = \frac{1-b}{a} \quad (77)$$

Comparing with Eq. (60) we find that $d_{n1} = d_{u1}$. This has arisen because of the equality of the indices of refraction on either side of the optical system. Similarly we can get

$$d_{n2} = \frac{c-1}{a} \quad (78)$$

Thus, when the media on either side of an optical system have the same refractive index (which is indeed the case for most optical systems), the nodal planes coincide with the unit planes. In general, if we know the elements of the system matrix S (i.e. if we know a , b , c and d which are also

called the Gaussian constants of the system), one can obtain all the properties of the system.

Example 4.2 Consider a thick equi-convex lens (made of a material of refractive index 1.5) of the type shown in Fig. 4.9. The magnitudes of the radii of curvature of the two surfaces is 4 cm. The thickness of the lens is 1 cm and the lens is placed in air. Obtain the system matrix and determine the focal length and the positions of unit planes.

Solution:

$$R_1 = +4 \text{ cm} \quad R_2 = -4 \text{ cm} \quad t = 1 \text{ cm}$$

Both surfaces have equal power

$$P_1 = P_2 = \frac{n-1}{R_1} = \frac{0.5}{4} = 0.125 \text{ cm}^{-1}$$

Thus the system matrix is from Eq. (51)

$$\begin{pmatrix} 1 - \frac{0.125 \times 1}{1.5} & -0.125 - 0.125 \left(1 - \frac{1}{1.5} \times 0.125 \right) \\ \frac{1}{1.5} & 1 - \frac{0.125}{1.5} \end{pmatrix} = \begin{pmatrix} 0.9167 & -0.240 \\ 0.6667 & 0.9167 \end{pmatrix}$$

Thus

$$a = \frac{1}{f} = 0.24 \Rightarrow f \approx 4.2 \text{ cm}$$

$$b = 0.9167 = c, \quad d = -0.6667$$

Using Eqs (60) and (61), we get the positions of the unit planes

$$d_{u1} = \frac{1-b}{a} \approx 0.35 \text{ cm}$$

$$d_{u2} = \frac{c-1}{a} \approx -0.35 \text{ cm}$$

Thus the unit planes are as shown in Fig. 4.9. The nodal planes coincide with the unit planes because the lens is immersed in air.

Example 4.3 Consider a sphere of radius 20 cm of refractive index 1.6 (see Fig. 4.11). Find the positions of the paraxial focal point and the unit planes.

Solutions: The matrices from the first refracting surface to the image plane are given by

Second surface to image	Refraction at second surface	Transmission through glass	Refraction at the first surface
$\begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & (1-1.6)/20 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 40/1.6 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & -(1.6-1)/20 \\ 0 & 1 \end{pmatrix}$

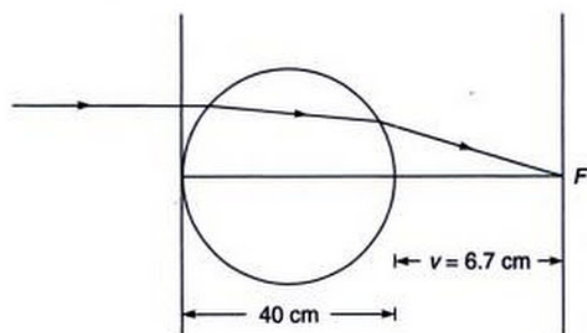


Fig. 4.11 Imaging by a sphere of radius 20 cm and refractive index 1.6.

$$= \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix} \begin{pmatrix} 0.25 & -0.0375 \\ 25 & 0.25 \end{pmatrix}$$

$$= \begin{pmatrix} 0.25 & -0.0375 \\ 25 + 0.25v & 0.25 - 0.0375v \end{pmatrix}$$

Thus at the image plane, the ray coordinates are

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.25 & -0.0375 \\ 25 + 0.25v & 0.25 - 0.0375v \end{pmatrix} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

This gives us

$$x_2 = (25 + 0.25v)\lambda_1 + (0.25 - 0.0375v)x_1$$

To determine the focal distance v , consider a ray incident parallel to the axis for which $\lambda_1 = 0$. The focal plane would be that plane for which x_2 is also zero. This gives us

$$0.0375v = 0.25 \quad \text{or} \quad v = 6.7 \text{ cm}$$

The system matrix elements are

$$a = \frac{1}{f} = 0.0375 \text{ cm}^{-1} \Rightarrow f \approx 26.7 \text{ cm}$$

$$b = 0.25, \quad c = 0.25, \quad d = -25 \text{ cm}$$

The unit planes are given by

$$d_{u1} = \frac{1-b}{a} = 20 \text{ cm}$$

and

$$d_{u2} = \frac{c-1}{a} = -20 \text{ cm}$$

Thus both the unit planes pass through the centre of the sphere.

4.5 A SYSTEM OF TWO THIN LENSES

We finally use the matrix formulation for the analysis of a combination of two thin lenses of focal lengths f_1 and f_2

separated by a distance t . The system matrix for the combination of the two lenses can be obtained by noting that the matrix of the two lenses are [see Eq. (57)]

$$\begin{pmatrix} 1 & -\frac{1}{f_1} \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} 1 & -\frac{1}{f_2} \\ 0 & 1 \end{pmatrix} \quad (79)$$

and the matrix for translation through a distance t (in air) is

$$\begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} \quad (80)$$

Thus the system matrix S is given by:

$$S = \begin{pmatrix} 1 & -\frac{1}{f_2} \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ t & 1 \end{pmatrix} \begin{pmatrix} 1 & -\frac{1}{f_1} \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \left(1 - \frac{t}{f_2}\right) & -\left(\frac{1}{f_1} + \frac{1}{f_2} - \frac{t}{f_1 f_2}\right) \\ t & \left(1 - \frac{t}{f_1}\right) \end{pmatrix} \quad (81)$$

Thus

$$\left. \begin{aligned} a &= \frac{1}{f_1} + \frac{1}{f_2} - \frac{t}{f_1 f_2}, & b &= 1 - \frac{t}{f_2} \\ c &= 1 - \frac{t}{f_1}, & d &= -t \end{aligned} \right\} \quad (82)$$

As already noted, the element a in the system matrix represents the inverse of the focal length of the system. Thus, the focal length of the combination is

$$\frac{1}{f} = \frac{1}{f_1} + \frac{1}{f_2} - \frac{t}{f_1 f_2} = a \quad (83)$$

The positions of the unit planes are given by [see Eqs. (60) and (61)]:

$$d_{u1} = \frac{1-b}{a} = \frac{tf}{f_2}$$

$$d_{u2} = \frac{c-1}{a} = -\frac{tf}{f_1} \quad (84)$$

It is easy to see that if we have a system of four thin lenses, we simply have to multiply seven matrices [four of them being of the type given by Eq. (79) and three of them of the type given by Eq. (80)].

Example 4.4 Consider a lens combination consisting of a convex lens (of focal length + 15 cm) and a concave lens (of focal length - 20 cm) separated by 25 cm (see Fig. 4.12 and Problem 3.5). Determine the system matrix elements and the positions of the unit planes. For an object (of height 1 cm) placed at a distance of 27.5 cm from the convex lens, determine the size and position of the image.

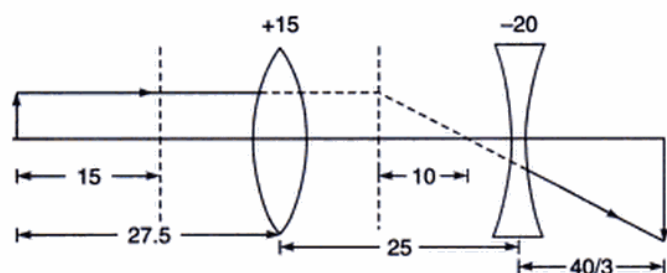


Fig. 4.12

Solution:

$$f_1 = +15 \text{ cm} \quad f_2 = -20 \text{ cm} \quad t = 25 \text{ cm}$$

Thus, using Eq. (82), we readily get

$$a = \frac{1}{10} = \frac{1}{f}, \quad b = \frac{45}{20}, \quad c = -\frac{2}{3}, \quad d = -25$$

and

$$d_{u1} = \frac{1-b}{a} = -12.5 \text{ cm}, \quad d_{u2} = \frac{c-1}{a} = -\frac{50}{3} \text{ cm}$$

Thus the distance of the object from the first unit plane is given by

$$u = -27.5 - (-12.5) = -15 \text{ cm}$$

Since $f = +10 \text{ cm}$, we get [using Eq. (67)]

$$v = 30 \text{ cm}$$

which represents the distance of the image plane from the second unit plane. Thus the image is at a distance of $30 - (50/3) = 40/3 \text{ cm}$ from the concave lens. The magnification is given by

$$M = \frac{v}{u} = -2$$

Example 4.5 Consider a system of two thin lenses as shown in Fig. 3.10. For a 1 cm tall object at a distance of 40 cm from the convex lens, calculate the position and size of the image.

Solution: Let v be the distance of the image plane from the concave lens. Thus the matrix, which when operated on the object column matrix gives the image column matrix, is given by

Concave lens to image	Concave lens	Convex lens to concave lens	Convex lens	Object to convex lens
$\begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & +1/10 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 8 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & -1/20 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 40 & 1 \end{pmatrix}$
$= \begin{pmatrix} 1 & 0 \\ v & 1 \end{pmatrix} \begin{pmatrix} 2.2 & 0.01 \\ +32 & 0.6 \end{pmatrix}$				

$$= \begin{pmatrix} 2.2 & 0.01 \\ 2.2v + 32 & 0.6 + 0.01v \end{pmatrix}$$

The image plane would correspond to

$$32 + 2.2v = 0$$

or

$$v \approx -14.5 \text{ cm}$$

i.e. it is at a distance of 14.5 cm to the left of the concave lens. If we compare this with Eq. (45), we obtain

$$M = 0.6 + 0.01v = 0.6 - 0.01 \frac{32}{2.2} = +\frac{1}{2.2}$$

Example 4.6 In the above example determine the system matrix and hence the positions of the unit planes. Finally, use Eq. (67) to determine the position of the image.

Solution: The system matrix is given by

$$S = \begin{pmatrix} 1 & 1/10 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 8 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1/20 \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 9/5 & 1/100 \\ 8 & 3/5 \end{pmatrix}$$

Thus

$$a = -\frac{1}{100} \Rightarrow f = -100 \text{ cm}$$

$$b = \frac{9}{5}, \quad c = \frac{3}{5}, \quad d = -8$$

If we now use Eqs (60 and (61), we have

$$d_{u1} = \frac{1-b}{a} = 80 \text{ cm}$$

and

$$d_{u2} = \frac{c-1}{a} = 40 \text{ cm}$$

Thus the first unit plane is at a distance of 80 cm to the right of the convex lens and the second unit plane is at 40 cm to the right of the concave lens. The object distance from the first unit plane is therefore given by

$$u = -(80 + 40) = -120 \text{ cm}$$

We now use Eq. (67) to obtain

$$\frac{1}{v} = a + \frac{1}{u} = -\frac{1}{100} - \frac{1}{120} = -\frac{22}{1200}$$

$$\Rightarrow v = -\frac{600}{11} \text{ cm}$$

Thus the image is at 54.5 cm to the left of the second unit plane or at 14.5 cm to the left of the concave lens as shown in Fig. 3.10. The magnification is

$$M = \frac{v}{u} = +\frac{1}{2.2}$$

SUMMARY

- In the paraxial approximation we may confine ourselves to rays which pass through the axis of the system; these rays remain confined to a single plane. Such a ray can be specified by its distance from the axis of the system x , and the quantity $\lambda = n \sin \alpha$ which represents the product of the refractive index with the sine of the angle that the ray makes with z -axis.
- If a ray is initially specified by a (2×1) matrix with elements λ_1 and x_1 , then the effect of translation through a distance D in a homogenous medium of refractive index n_1 , is given by

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = T \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

where the translation matrix T is given by

$$T = \begin{pmatrix} 1 & 0 \\ D/n_1 & 1 \end{pmatrix}$$

- The effect of refraction through a spherical refracting surface (separating media of refractive indices n_1 and n_2) is given by

$$\begin{pmatrix} \lambda_2 \\ x_2 \end{pmatrix} = \mathcal{R} \begin{pmatrix} \lambda_1 \\ x_1 \end{pmatrix}$$

where the refraction matrix is given by

$$\mathcal{R} = \begin{pmatrix} 1 & -P \\ 0 & 1 \end{pmatrix}$$

with

$$P = \frac{n_2 - n_1}{R}$$

- By successive application of the above matrices one can study paraxial imaging by a coaxial optical system.
- In an optical system, unit planes are two planes, one each in the object and the image space, between which the magnification M is unity, i.e., any paraxial ray emanating from the unit plane in the object space will emerge at the same height from the unit plane in the image space.
- Nodal points are two points on the axis which have a relative angular magnification of unity, i.e., a ray striking the first point at an angle α emerges from the second point at the same angle. The planes which pass through these points and are normal to the axis are known as nodal planes.

PROBLEMS

- 4.1 Consider a system of two thin convex lenses of focal lengths 10 and 30 cm separated by a distance of 20 cm in air.

- Determine the system matrix elements and the positions of the unit planes.
- Assume a parallel beam of light incident from the left. Use Eq. (67) and the positions of the unit planes to determine the image point. Using the unit planes draw the ray diagram.

[Ans: (a) $a = 1/15$, $b = 1/3$, $c = -1$, $d = -20$; the first convex lens is in the middle of the two unit planes. (b) The final image is virtual and is 15 cm away (on the left) from the second lens.]

- 4.2 Consider a thick biconvex lens whose magnitudes of the radii of curvature of the first and second surfaces are 45 cm and 30 cm respectively. The thickness of the lens is 5 cm and the refractive index of the material of the lens is 1.5. Determine the elements of the system matrix and positions of the unit planes and use Eq. (67) to determine the image point of an object at a distance of 90 cm from the first surface.

[Ans: $a = 0.02716$, $b = 0.9444$, $c = 0.9630$, $d = -3.3333$, $d_{u1} = 2.0455$, $d_{u2} = -1.3636$. Final image at a distance of 60 cm from the second surface.]

- 4.3 Consider a hemisphere of radius 20 cm and refractive index 1.5. If H_1 and H_2 denote the positions of the first and second principal points, then show that $AH_1 = 13.3$ cm and that H_2 lies on the second surface as shown in Fig. 4.13. Further, show that the focal length is 40 cm.

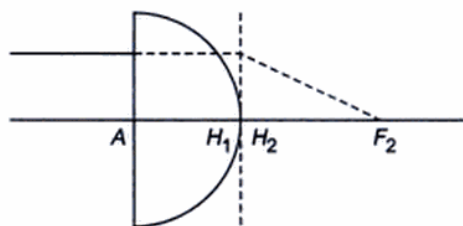


Fig. 4.13

- 4.4 Consider a thick lens of the form shown in Fig. 4.14; the radii of curvature of the first and second surfaces are -10 cm and $+20$ cm respectively and the thickness of the lens is 1.0 cm. The refractive index of the material of the lens is 1.5 . Determine the positions of the principal planes.

[Ans: $d_{u1} = 20/91$ cm, $d_{u2} = 40/91$ cm]

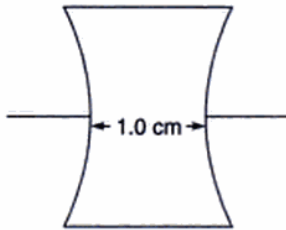


Fig. 4.14

- 4.5 Consider a combination of two thin lenses of focal lengths f_1 and f_2 separated by a distance $(f_1 + f_2)$. Show that the angular magnification of the lens

combinations (which is just $\frac{\lambda_2}{\lambda_1} = \frac{\alpha_2}{\alpha_1}$) is given by

$-f_1/f_2$. Interpret the negative sign in the expression for magnification.

- 4.6 Consider a spherical refracting surface as shown in Fig. 3.12. Using matrix method show that for an

object at a distance of $\left(1 + \frac{n_2}{n_1}\right)r$ from the surface,

the image is virtual and at a distance of $\left(1 + \frac{n_1}{n_2}\right)r$ from the surface.

REFERENCES AND SUGGESTED READINGS

1. J.N. Blaker, *Geometrical Optics: The Matrix Theory*, Marcel Dekker, New York, 1971.
2. W. Brouwer, *Matrix Methods in Optical Instrumental Design*, Benjamin, New York, 1964.
3. D.M. Eakin, and S.P. Davis, 'An application of matrix optics,' *American Journal of Physics*, Vol. 34, 758, 1966.
4. A. Gerrard, and J.M. Burch, *Introduction to Matrix Methods in Optics*, John Wiley & Sons, New York, 1975.
5. K. Halbach, 'Matrix representation of Gaussian optics,' *American Journal of Physics*, Vol. 32, 90, 1964.
6. A Nussbaum, *Geometric Optics: An Introduction*, Addison-Wesley Publishing Co., Reading, Mass, 1968.
7. J.W. Simmons and M.J. Guttman, *States, Waves and Photons: A Modern Introduction to Light*, Addison-Wesley Publishing Co., Reading, Mass, 1970.

Chapter 5

Aberrations

Geometrical optics is either very simple or else it is very complicated..... If one has an actual, detailed problem in lens design, including analysis of aberrations, then he has to simply trace the rays through the various surfaces using the law of refraction and find out where they come out and see if they form a satisfactory image. People have said that this is too tedious, but today, with computing machines, it is the right way to do it. One can set up the problem and make the calculation one ray after another very easily. So the subject is really ultimately quite simple, and involves no new principles.

— Richard Feynman in *Feynman Lectures in Physics, Vol. I*

5.1 INTRODUCTION

In Chapter 3, while studying the formation of images by refracting surfaces and thin lenses, we had made the assumption that the object point does not lie far away from the axis of the optical system and that the rays taking part in image formation are essentially those which make small angles with the axis of the system. In practice, neither of the above assumptions is true; one in fact has to deal with rays making large angles with the axis. The domain of optics dealing with rays lying close to the optical axis and making small angles with it is called paraxial optics. We had found that in the realm of paraxial optics, the images of objects were perfect, i.e. all rays emanating from a single object point converged to a single image point and the magnification of the system was a constant of the optical system, independent of the particular ray under consideration. Since in real optical systems, nonparaxial rays also take part in image formation, the actual images depart from the ideal images. This departure leads to what are known as aberrations.

It can be shown that the primary aberrations of any rotationally symmetric system can be specified by five coefficients. The five coefficients represent the spherical aberration, coma, astigmatism, curvature of field and distortion. These are called the Seidel aberrations. Since these aberrations are present even for light of a single wavelength, they are also called monochromatic aberrations. In this chapter, we will consider the five kinds of aberrations separately and discuss the effect on the image when each one of them is present separately.

It should be mentioned that if a polychromatic source (like white light) is used for image formation (which is

indeed the case for many optical instruments) then, in general, the images will be coloured; this is known as chromatic aberration. Physically, chromatic aberration is due to the dependence of the refractive index of the material of the lens on wavelength of the radiation under consideration. Since images formation is accompanied by refraction at refractive index discontinuities, the wavelength dependence of the refractive index results in the coloured image. For a polychromatic source, different wavelength components (after refraction) proceed along different directions and form images at different points; this leads to coloured images. Since chromatic aberration is the easiest to understand, we would discuss this first. This will be followed by a discussion of monochromatic aberrations.

5.2 CHROMATIC ABERRATION

Let us consider a parallel beam of white light incident on a thin convex lens as shown in Fig. 5.1. Since blue light gets refracted more than red light, the point at which the blue light would focus is nearer the lens than the point at which the red light would focus. Thus, the image will appear to be coloured; it may be mentioned that this aberration is independent of the five Seidel aberrations to be discussed in later sections.

For the case of a thin lens, the expression for chromatic aberration can easily be derived. The focal length of a thin lens is given by

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (1)$$

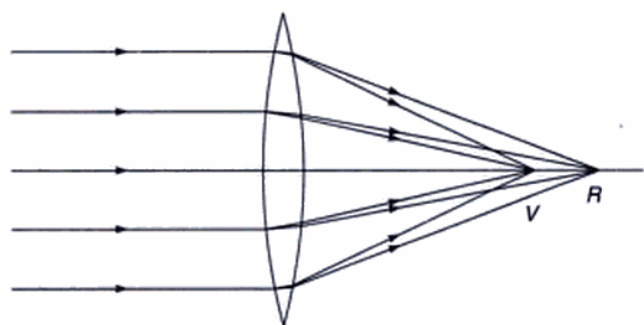


Fig. 5.1 When white light consisting of a continuous range of wavelengths is incident on a lens, then each wavelength refracts by different amounts; this leads to chromatic aberration. This aberration is independent of five Seidel aberrations.

If a change of n by δn (the change of n is due to the change in the wavelength of the light) results in a change of f by δf then we obtain by differentiating Eq. (1)

$$-\frac{\delta f}{f^2} = \delta n \left(\frac{1}{R_1} - \frac{1}{R_2} \right) = \frac{\delta n}{n-1} \frac{1}{f}$$

$$\text{i.e.,} \quad \delta f = -f \frac{\delta n}{n-1} \quad (2)$$

which represents the chromatic aberration of a thin lens. If n_b and n_r represent the refractive indices for the blue and red colours respectively, then

$$f_r - f_b = f \left(\frac{n_b - n_r}{n-1} \right) \quad (3)$$

would represent the chromatic aberration.

5.2.1 The Achromatic Doublet

We will first consider an optical system of two thin lenses made of different materials placed in contact with each other. For example, one of the lenses may be made of crown glass and the other of flint glass. We will find the condition for this lens combination to have the same focal length for the blue and red colours. Let n_b , n_y and n_r represent the refractive indices for the material of the first lens corresponding to the blue, yellow and red colours respectively. Similarly, n'_b , n'_y and n'_r represent the corresponding refractive indices for the second lens. If f_b and f'_b represent the focal lengths for the first and the second lens corresponding to the blue colour, and if F_b represents the focal length of the combination of the two lenses (placed in contact), then

$$\frac{1}{F_b} = \frac{1}{f_b} + \frac{1}{f'_b} = (n_b - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) + (n'_b - 1) \left(\frac{1}{R'_1} - \frac{1}{R'_2} \right) \quad (4)$$

where R_1 and R_2 represent the radii of curvatures of the first and second surface for the first lens and, as before, the primed quantities refer to the second lens. Thus, we may write

$$\frac{1}{F_b} = \frac{n_b - 1}{n - 1} \frac{1}{f} + \frac{n'_b - 1}{n' - 1} \frac{1}{f'} \quad (5)$$

where

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right),$$

$$\frac{1}{f'} = (n' - 1) \left(\frac{1}{R'_1} - \frac{1}{R'_2} \right) \quad (6)$$

$$n = \frac{n_b + n_r}{2} \approx n_y, \quad n' = \frac{n'_b + n'_r}{2} \approx n'_y \quad (7)$$

f and f' represent the focal lengths of the first and second lens corresponding to a mean colour which is around the yellow region. Similarly, the focal length of the combination corresponding to the red colour would be given by

$$\frac{1}{F_r} = \frac{n_r - 1}{n - 1} \frac{1}{f} + \frac{n'_r - 1}{n' - 1} \frac{1}{f'} \quad (8)$$

For the focal length of the combination to be equal for blue and red colours, we must have

$$\frac{n_b - 1}{n - 1} \frac{1}{f} + \frac{n'_b - 1}{n' - 1} \frac{1}{f'} = \frac{n_r - 1}{n - 1} \frac{1}{f} + \frac{n'_r - 1}{n' - 1} \frac{1}{f'}$$

or

$$\frac{\omega}{f} + \frac{\omega'}{f'} = 0 \quad (9)$$

where

$$\omega = \frac{n_b - n_r}{n - 1} \quad \text{and} \quad \omega' = \frac{n'_b - n'_r}{n' - 1} \quad (10)$$

are known as the dispersive powers. Since ω and ω' are both positive, f and f' must be of opposite signs for the validity of Eq. (9). A lens combination which satisfies Eq. (9) is known as an achromatic doublet (see Fig. 5.2). It may be mentioned that if the two lenses are made of the same material, then $\omega = \omega'$ and Eq. (9) would imply $f = -f'$; such a combination will have an infinite focal length. Thus, for an achromatic doublet the two lenses must be of different materials.

Example 5.1 An achromatic doublet of focal length 20 cm is to be made by placing a convex lens of borosilicate crown glass in contact with a diverging lens of dense flint glass. Assuming $n_r = 1.51462$, $n_b = 1.52264$, $n'_r = 1.61216$ and $n'_b = 1.62901$, calculate the focal length of each lens; here the unprimed and the primed quantities refer to the borosilicate crown glass and dense flint glass respectively.

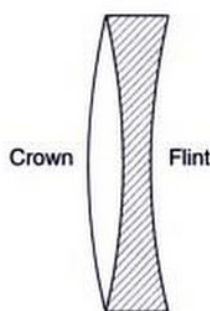


Fig. 5.2 An achromatic doublet.

Solution:

$$n \approx \frac{n_b + n_r}{2} = \frac{1.52264 + 1.51462}{2} = 1.51863$$

$$n' \approx \frac{n'_b + n'_r}{2} = \frac{1.62901 + 1.61216}{2} = 1.62058$$

Thus,

$$\omega = \frac{1.52264 - 1.51462}{1.51863 - 1} = 0.01546$$

and

$$\omega' = \frac{1.62901 - 1.61216}{1.62058 - 1} = 0.02715$$

Substituting in Eq. (9), we obtain

$$\frac{0.01546}{f} + \frac{0.02715}{f'} = 0$$

$$\text{or} \quad \frac{f}{f'} = -0.56942$$

Now, for the lens combination to be of focal length 20 cm we must have

$$\frac{1}{f} + \frac{1}{f'} = \frac{1}{20}$$

or

$$\frac{1}{f} [1 - 0.56942] = \frac{1}{20}$$

$$\text{or} \quad f = 20 \times 0.43058 = 8.61 \text{ cm}$$

and

$$f' = -\frac{f}{0.56942} \approx -15.1 \text{ cm}$$

5.2.2 Removal of Chromatic Aberration of a Separated Doublet

Let us consider two thin lenses of focal lengths f and f' and separated by a distance t (see Fig. 5.3). The focal length of the combination F , would be

$$\frac{1}{F} = \frac{1}{f} + \frac{1}{f'} - \frac{t}{ff'} \quad (11)$$

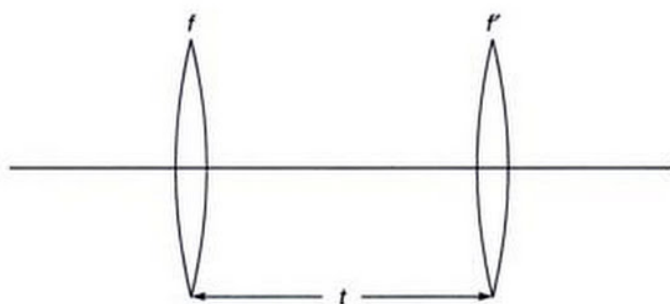


Fig. 5.3 The separated doublet.

The focal length of the first lens would be given by

$$\frac{1}{f} = (n - 1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (12)$$

with a similar expression for $1/f'$. If Δf and Δn represent the changes in the focal length and in the refractive index due to a change $\Delta \lambda$ in the wavelength, then by differentiating Eq. (12), we obtain

$$-\frac{\Delta f}{f^2} = \Delta n \left(\frac{1}{R_1} - \frac{1}{R_2} \right) = \frac{\Delta n}{(n - 1)f}$$

Thus, differentiating Eq. (11), we obtain

$$\begin{aligned} -\frac{\Delta F}{F^2} &= -\frac{\Delta f}{f^2} - \frac{\Delta f'}{f'^2} + \frac{t}{f} \frac{\Delta f'}{f'^2} + \frac{t}{f'} \frac{\Delta f}{f^2} \\ &= \frac{\Delta n}{(n - 1)f} + \frac{\Delta n'}{(n' - 1)f'} - \frac{t}{f} \frac{\Delta n'}{(n' - 1)f'} - \frac{t}{f'} \frac{\Delta n}{(n - 1)f} \\ &= \frac{\omega}{f} + \frac{\omega'}{f'} - \frac{t}{ff'} (\omega + \omega') \end{aligned} \quad (13)$$

where, as before, ω and ω' represent the dispersive powers. Consequently, for the combination to have the same focal length for blue and red colours we should have

$$\frac{t(\omega + \omega')}{ff'} = \frac{\omega}{f} + \frac{\omega'}{f'}$$

or

$$t = \frac{\omega f' + \omega' f}{\omega + \omega'} \quad (14)$$

If both the lenses are made of the same material, then $\omega = \omega'$ and the above equation simplifies to

$$t = \frac{f + f'}{2} \quad (15)$$

implying that the chromatic aberration is very small if the distance between the two lenses is equal to the mean of the focal lengths. This is indeed the case for the Huygens' eyepiece.

5.3 MONOCHROMATIC ABERRATIONS

5.3.1 Spherical Aberration

Let a beam of light parallel to the axis be incident on a thin lens (see Fig. 5.4). The light rays after passing the lens bend towards the axis and cross the axis at some point. If we restrict ourselves to the paraxial region, then we can see that all rays cross the z -axis at the same point which is at a distance f_p from the lens; f_p represents the paraxial focal length of the lens. If one does not restrict to the paraxial region, then in general, rays which are incident at different heights on the lens, hit the axis at different points. For example, for a convex lens, the marginal rays (which are incident near the periphery of the lens) focus at a point closer than the focal point of paraxial rays [see Fig. 5.4(a)]. Similarly, for a concave lens, rays which are incident farther from the axis appear to be emerging from a point which is nearer to the lens [see Fig. 5.4(b)]. The point at which the paraxial rays strike the axis (F_P) is called the paraxial focus and the point at which the rays near the periphery strike is called the marginal focus (F_M). The distance between the two foci is a measure of spherical aberration in the lens. Thus if O represents an axial object, then different rays emerging from the object converge to different points; consequently, the image of a point object will not be a point. The distance along the axis between the paraxial image point and the image corresponding to marginal rays (i.e., rays striking the edge of the lens) is termed longitudinal spherical aberration. Similarly, the distance between the paraxial image point and the point at which the marginal ray strikes the paraxial image plane is called the lateral spherical aberration [see Fig. 5.4(a)]. The image on any plane (normal to the z -axis) is a circular patch of light; however, as can be seen from Fig. 5.4(a), on a plane AB the circular patch has the least diameter. This is called the **circle of least confusion** (see Fig. 5.5). It may be mentioned that for an object lying on the axis of a cylindrically symmetric system (like a system of coaxial lenses), the image will suffer only from spherical aberration. All other off-axis aberrations like coma, astigmatism, etc., will be absent.

To see how the rays hitting the refracting surface at different heights could focus to different points on the axis, let us consider the simple case of a plane refracting surface as shown in Fig. 5.6. Let the plane of the refracting surface be chosen as the plane $z = 0$. Let P be the object point. The z -axis is chosen to be along the normal (PO) from the point P to the surface. The plane $z = 0$ separates two media of refractive indices n_1 and n_2 (see Fig. 5.6); in the figure we have assumed $n_2 > n_1$. Consider a ray PM incident on the

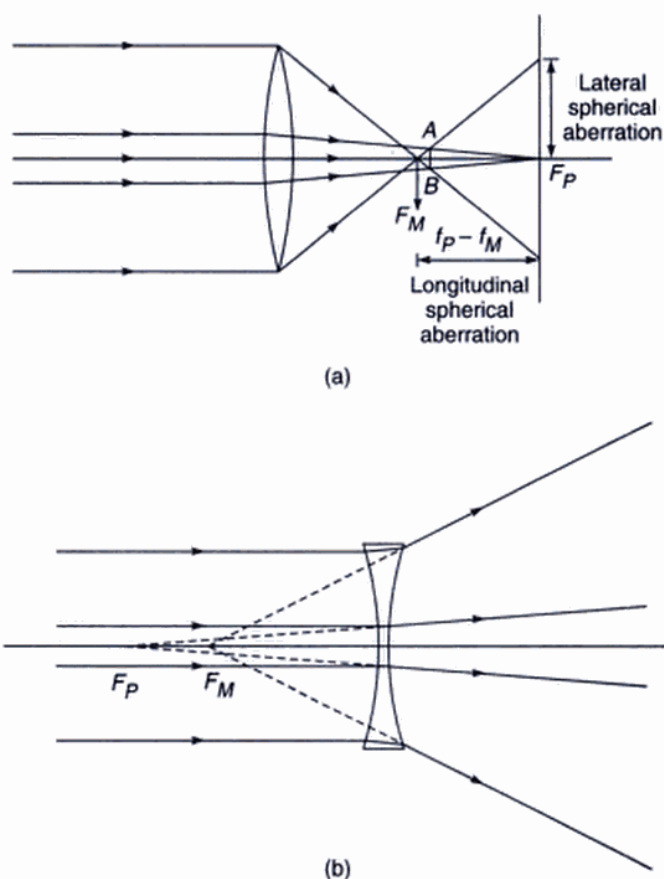


Fig. 5.4 (a) For a converging lens the focal point for marginal rays lies closer to the lens than the focal point for paraxial rays. The distance between the paraxial focal point and the marginal focal point is known as the longitudinal spherical aberration and the radius of the image at the paraxial focal plane is known as the lateral spherical aberration. The combined effect of defocusing and spherical aberration leads to the formation of a circle of least confusion, where the image would have the minimum diameter. (b) The spherical aberration of a diverging lens.

refracting surface (from the object) at a height h as shown in Fig. 5.6. The refracted ray appears to emerge from the point Q . We assume the origin to be at the point O . Let the z -coordinates of the points P and Q be z_0 and z_1 respectively. Obviously, both z_0 and z_1 would be negative quantities and the distances OP and OQ would be $-z_0$ and $-z_1$ respectively (see Fig. 5.6). We have to determine z_1 in terms of z_0 . From Snell's law we know that

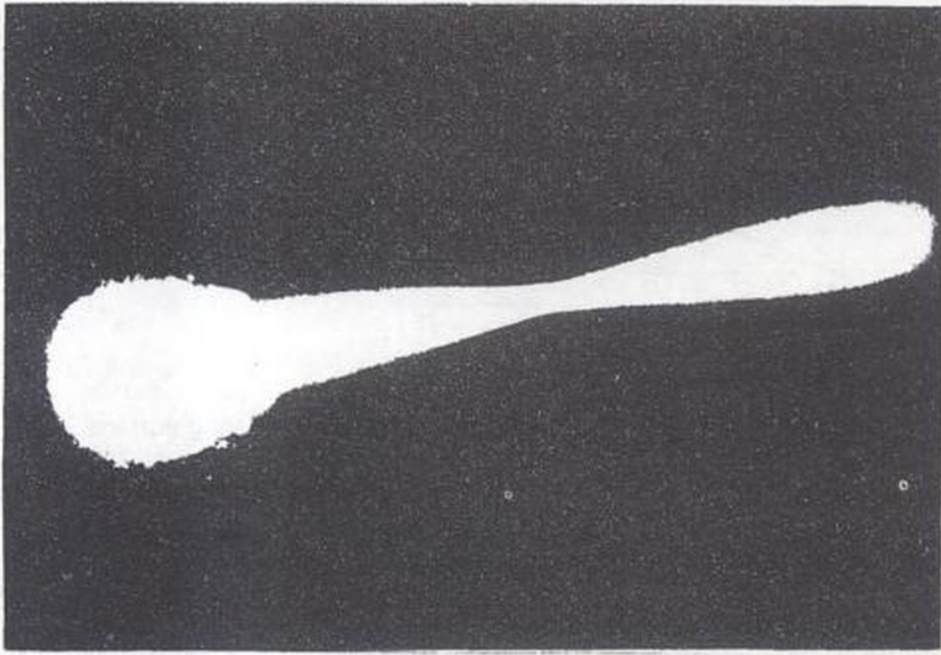


Fig. 5.5 The spherical aberration of a convex lens (photograph courtesy Dr. K.K. Gupta).

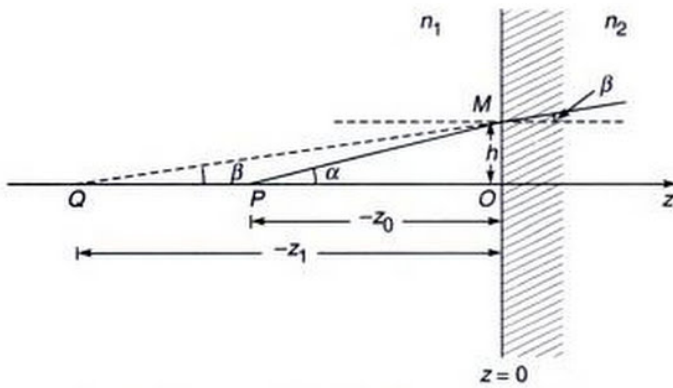


Fig. 5.6 Refraction at a plane surface.

$$\sin \alpha = n \sin \beta \quad (16)$$

where α and β are the angles that the incident and refracted rays make with the z -axis and

$$n = \frac{n_2}{n_1} \quad (17)$$

Now, from Fig. 5.6 we have

$$(-z_1) = h \cot \beta = \frac{h}{\sin \beta} \sqrt{1 - \sin^2 \beta}$$

or

$$z_1 = - \frac{nh}{\sin \alpha} \left(1 - \frac{1}{n^2} \sin^2 \alpha \right)^{1/2} \quad (18)$$

where we have used Eq. (16). Since

$$\sin \alpha = \frac{h}{\sqrt{h^2 + z_0^2}} \quad (19)$$

we obtain

$$z_1 = - \frac{nh}{h} (h^2 + z_0^2)^{1/2} \left[1 - \frac{1}{n^2} \frac{h^2}{(h^2 + z_0^2)} \right]^{1/2} \quad (20)$$

$$z_1 = -n |z_0| \left[1 + \frac{h^2}{z_0^2} \right]^{1/2} \left[1 - \frac{h^2}{n^2 z_0^2} \left(1 + \frac{h^2}{z_0^2} \right)^{-1} \right]^{1/2} \quad (21)$$

The value of z_1 given in Eq. (21) is an exact expression in terms of z_0 . It can at once be seen that the image distance, z_1 , is a complicated function of the height h , at which the ray strikes the refracting surface. In the limit of $h \rightarrow 0$, i.e. for paraxial rays, we get

$$z_1 = -n |z_0| \quad (22)$$

which is the expression for the image distance in the paraxial region. To the next order of approximation, assuming $|h/z_0| \ll 1$, we get

$$\begin{aligned} z &\approx -n |z_0| \left[1 + \frac{h^2}{2 z_0^2} \right] \left[1 - \frac{h^2}{2 n^2 z_0^2} \right] \\ &\approx -n |z_0| \left[1 + \frac{h^2}{2 z_0^2} (n^2 - 1) \right] \end{aligned} \quad (23)$$

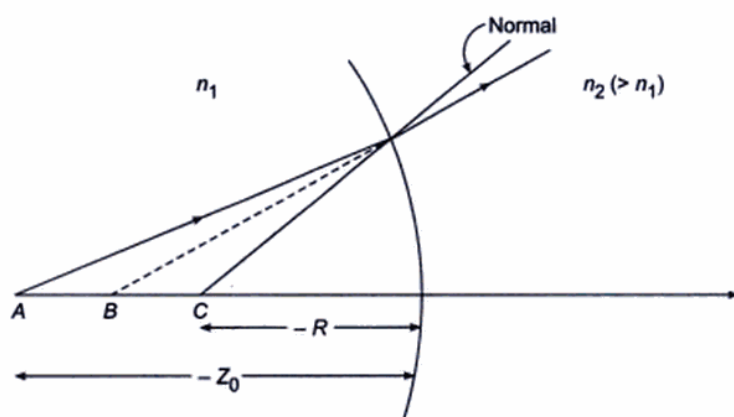


Fig. 5.7 The aplanatic points of a spherical refracting surface.

Thus the aberration is given by

$$\Delta z = -\frac{h^2}{2n|z_0|} (n^2 - 1) \quad (24)$$

Equation (24) gives the longitudinal spherical aberration. The negative sign implies that the non-paraxial rays appear to emanate from a point which is farther away from the paraxial image point.

From the above example, it can be seen that even a single plane refracting surface suffers from spherical aberration. Thus, spherical refracting surfaces and thin lenses must also suffer from spherical aberration.

The calculation of the spherical aberration even for a single spherical refracting surface is quite cumbersome (see, e.g. Ref. 5); we just give the final results:

$$\Delta z = -\frac{(n_2 - n_1)}{2n_2 \left(\frac{1}{z_0} + \frac{n_2 - n_1}{n_1 R} \right)^2} \left(\frac{1}{R} + \frac{1}{z_0} \right)^2 \times \left(-\frac{n_2 + n_1}{n_1 z_0} + \frac{1}{R} \right) h^2 \quad (25)$$

where R represents the radius of curvature of the surface, n_1 and n_2 represent the refractive indices of the media on the left and right of the spherical surface (see Fig. 5.7). For a plane surface $R = \infty$, Eq. (25) reduces to Eq. (24) with $n = n_2/n_1$.

Example 5.2 Consider a spherical refracting surface of radius R . Show that for a point A [see Fig. 5.7(b)] such that

$$z_0 = \frac{n_1 + n_2}{n_1} R \quad (26)$$

the spherical aberration is zero. Notice that both R and z_0 are negative quantities. The corresponding image point B is

at a distance $\frac{n_2 - n_1}{n_2} z_0$. The points A and B are known as

the aplanatic points and are utilized in microscope objectives.

Solution: For $z_0 = \frac{n_1 + n_2}{n_1} R$, one of the factors in Eq. (25) vanishes and the spherical aberration is zero. Indeed, it can be rigorously shown that all rays emanating from the point A appear to diverge from the point B (see also Sec. 3.8).

Example 5.3 Consider a refracting surface obtained by revolving an ellipse about its major axis. Show that all the rays parallel to the major axis will focus at one of the foci if the eccentricity of the ellipse is equal to n_1/n_2 .

[Hint: The eccentricity of the ellipse is given by

$$\epsilon = \frac{OF}{a} = \left(1 - \frac{b^2}{a^2} \right)^{1/2}$$

where a and b are the semi-major and semi-minor axes respectively (see Fig. 5.8). If we assume $n_1(QP) + n_2(PF) = n_2(BF)$, then show that the coordinates of the point $P(x, y)$ will satisfy the equation of the ellipse.]

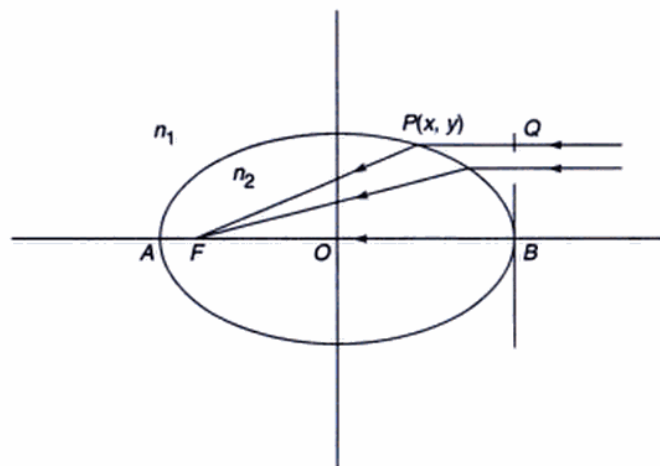


Fig. 5.8 For Example 5.3.

In a similar manner, for a set of rays incident parallel to the axis, one can show that the coefficient of spherical aberration of a thin lens made of a material of refractive index n and placed in air, with the surfaces having radii of curvatures R_1 and R_2 , would be given by

$$A = -\frac{f(n-1)}{2n^2} \times \left[-\left(\frac{1}{R_2} - P\right)^2 \left\{ \frac{1}{R_2} - P(n+1) \right\} + \frac{1}{R_1^3} \right] \quad (27)$$

where

$$P = \frac{1}{f} = (n-1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (28)$$

represents the power of the lens. The coefficient A is such that when it is multiplied by the cube of the height of the ray at the lens, one obtains the lateral spherical aberration. Thus the lateral spherical aberration for rays hitting the lens at a height h would be

$$S_{lat} = Ah^3 = -\frac{f(n-1)h^3}{2n^2} \times \left[-\left(\frac{1}{R_2} - P\right)^2 \left\{ \frac{1}{R_2} - P(n+1) \right\} + \frac{1}{R_1^3} \right] \quad (29)$$

The longitudinal spherical aberration which corresponds to the difference between the marginal focal length and the paraxial focal length would be given by

$$\begin{aligned} S_{long} &= Ah^2f \\ &= -\frac{(n-1)f^2h^2}{2n^2} \times \left[\frac{1}{R_1^3} - \left(\frac{1}{R_2} - \frac{n+1}{f} \right) \left(\frac{1}{R_2} - \frac{1}{f} \right)^2 \right] \end{aligned} \quad (30)$$

For a converging lens, S_{long} will always be negative implying that the marginal rays focus closer to the lens.

For a thin lens of given power (i.e. of a given focal length), one can define a quantity q , called the shape factor, by the following relation:

$$q = \frac{R_2 + R_1}{R_2 - R_1} \quad (31)$$

where R_1 and R_2 are the radii of curvatures of the two surfaces. For a given focal length of the lens, one can control the spherical aberration by changing the value of q . This procedure is called bending of the lens. Figure 5.9 shows the variation of spherical aberration with q for $n = 1.5$, $f = 40$ cm (i.e. $P = 0.025 \text{ cm}^{-1}$) and $h = 1$ cm. It can

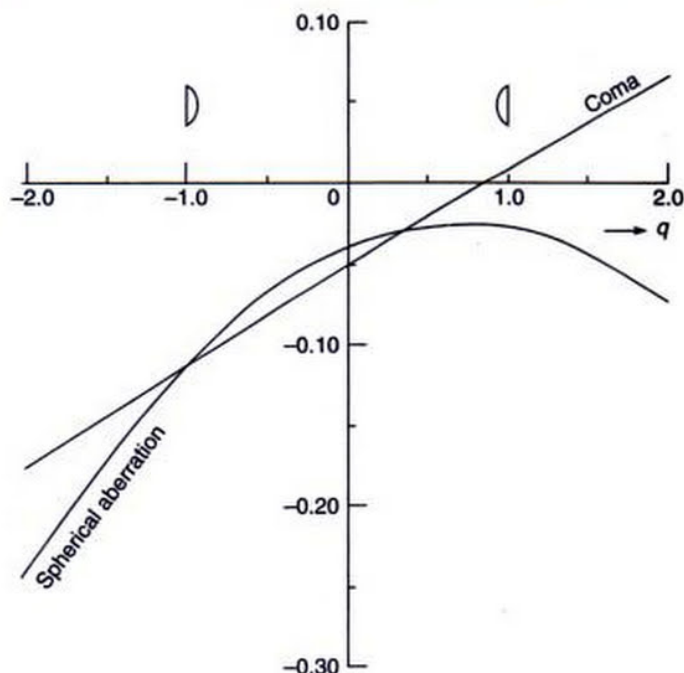


Fig. 5.9 Variation of spherical aberration and coma with the shape factor of a thin lens with $n = 1.5$, $f = 40$ cm and $h = 1$ cm. For calculating the coma we have assumed $\tan \theta = 1$, i.e., rays make an angle of 45° with the axis.

be seen that for values of q lying near $q \approx +0.7$, the (magnitude of the) spherical aberration is minimum (but not zero). Thus, by choosing proper values of the radii, the spherical aberration can be minimized. It may be mentioned that the value $q = +1$ implies $R_2 = \infty$ and hence it corresponds to a plano-convex lens with the convex side facing the incident light. On the other hand, for a plano-convex lens with the plane side facing the incident light $R_1 = \infty$ and $q = -1$. Thus the spherical aberration is dependent on how the deviation is divided between the surfaces.

The physical reason for the minimum of $|S_{long}|$ to occur at $q \approx 0.7$ is as follows: It has already been mentioned before that (for a converging lens) the marginal rays undergo a large deviation which results in the spherical aberration [see Fig. 5.4 (a)]. As such we should expect the spherical aberration to be minimum when the angle of deviation δ [see Fig. 5.10 (a)] is minimum. As in the case of the prism [see Fig. 5.10 (b)], this would occur when the deviations suffered at each of the refracting surfaces are exactly equal, i.e.,

$$\delta_1 = \delta_2; (\delta = \delta_1 + \delta_2) \quad (32)$$

Indeed for $q = 0.7$, the deviations suffered at each of the surface are equal and one obtains minimum spherical aberration.

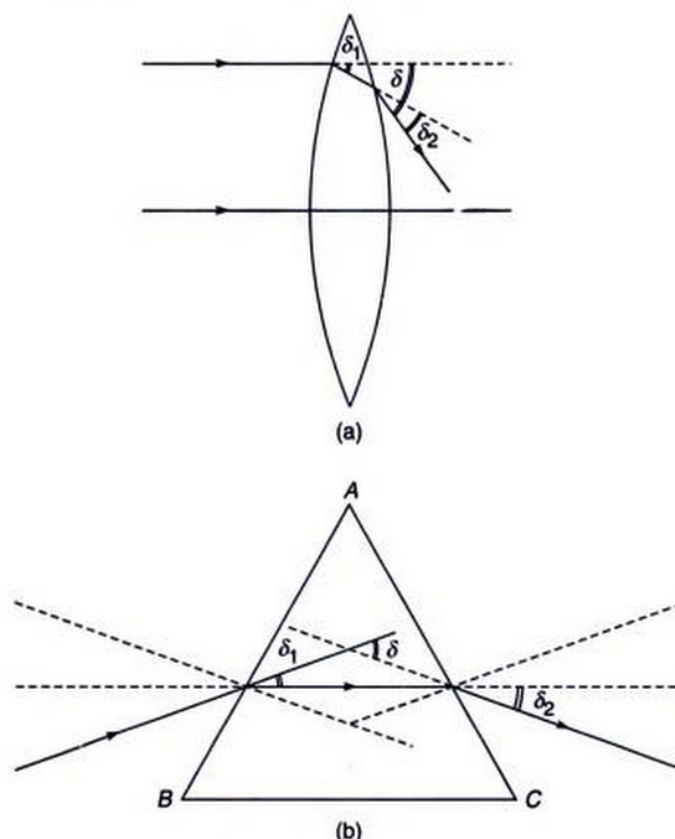


Fig. 5.10 (a) Refraction at the two refracting surfaces of a thin lens; the diagram is exaggerated to show clearly the angles. (b) For a prism, the minimum deviation position corresponds to $\delta_1 = \delta_2$.

Using the criterion of equal deviation discussed above, we will determine the separation between two thin lenses which would lead to minimum spherical aberration. Let L_1 and L_2 be two lenses of focal lengths f_1 and f_2 respectively separated by a distance x (see Fig. 5.11). If θ_1 and θ_2 represent the deviations of the ray at the two lenses, then for minimum spherical aberration, we get

$$\theta_1 = \theta_2 \quad (33)$$

To obtain an expression for the deviation suffered by a ray when it encounters a lens, we refer to Fig. 5.12 where a ray PA gets refracted along AQ after suffering a deviation through an angle θ . From triangle PAQ , we can see that

$$\begin{aligned} \theta &= \theta_1 + \theta_2 = \frac{h}{v} + \frac{h}{(-u)} \\ &= h \left(\frac{1}{v} - \frac{1}{u} \right) = \frac{h}{f} \end{aligned} \quad (34)$$

where we have used the paraxial relation

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \quad (35)$$

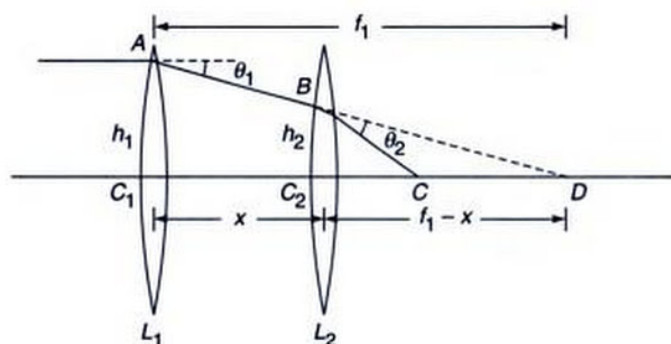


Fig. 5.11 Condition for minimum spherical aberration for a combination of two thin lenses.

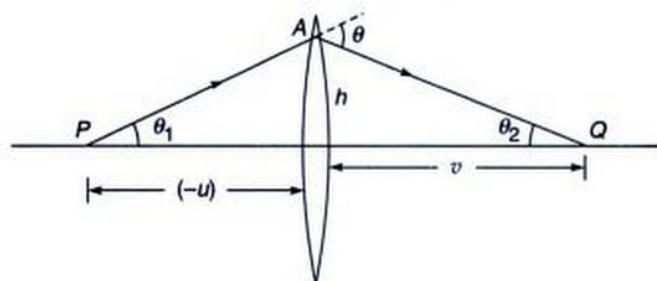


Fig. 5.12 Calculation of the angle of deviation.

The quantity u is an intrinsically negative quantity. Thus Eq. (33) becomes

$$\frac{h_1}{f_1} = \frac{h_2}{f_2} \quad (36)$$

From similar triangles AC_1D and BC_2D (see Fig. 5.11), we can write

$$\frac{h_1}{f_1} = \frac{h_2}{f_1 - x} \quad (37)$$

If we use Eqs. (36) and (37), we obtain

$$x = f_1 - f_2 \quad (38)$$

Thus the spherical aberration of a combination of two thin lenses is a minimum when their separation is equal to the difference in their focal lengths. Indeed, in the Huygens eyepiece (see Fig. 5.13), the focal length of the field lens is $3f$ where f represents the focal length of the eye lens. The distance between the two lenses is $2f$. We can immediately see that the conditions for achromatism [see Eq. (15)] and minimum spherical aberration [see Eq. (38)] are simultaneously satisfied. Since the eyepiece as a whole is corrected and the individual lenses are not, the image of the cross wires (which are placed in the plane PQ) will show aberrations. A discussion of the procedure for reducing the aberrations in various optical instruments requires a very detailed analysis involving the tracing of the rays, which is beyond the scope of this book.

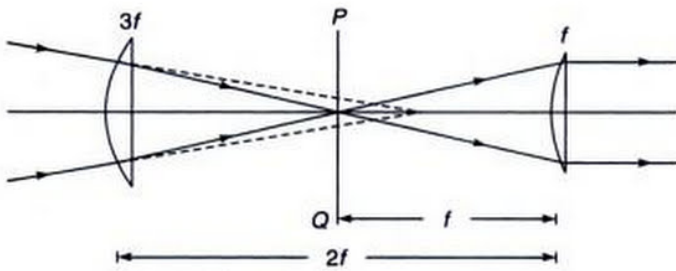


Fig. 5.13 The Huygens eyepiece.

It should be mentioned that even when the system is free from all aberrations the image of a point object will still not be a point because of diffraction effects (see Sec. 16.3). For example, if a perfectly spherical wave is emanating from a lens, the ray theory predicts a point image whereas the diffraction theory (which takes into account the finiteness of the wavelength) predicts that the image formed in the image plane will be an Airy pattern [see Fig. 16.8 (c)], and

the first dark ring will occur at a distance of $\frac{1.22\lambda f}{D}$ from

the paraxial image point (see Fig. 5.14) where D is the diameter of the exit pupil. The Airy pattern shown in Fig. 5.14 is highly magnified. For example, for $\lambda = 5000 \text{ \AA}$, $D = 5 \text{ cm}$, $f = 10 \text{ cm}$, the radii of the first and second dark rings in the Airy pattern will be about 0.00012 mm and 0.00022 mm respectively (see Sec. 16.3). The spatial extent of the Airy pattern will become larger with decrease in the value of D . Often one uses a 'stop' to restrict to the paraxial region; however, if the diameter of the 'stop' is made very small then the diffraction effects would dominate. Indeed, a

camera gives best image when $f/D \approx 5.6$; at high apertures aberrations degrade the image and at low apertures diffraction degrades the image.

5.3.2 Coma

As mentioned earlier for a point object lying on the axis the image will suffer only from spherical aberration. For off-axis points, the image will also suffer from coma, astigmatism, curvature of field and distortion. The first off-axis aberration is coma; i.e., for points lying very close to the axis, the image will suffer from spherical aberration and coma only. In this section we will briefly discuss the effect of coma, assuming that all other aberrations are absent.

The effect of coma is schematically shown in Fig. 5.15(a). The rays which proceed near the axis of the lens focus at a point different from that of the marginal rays. Thus, it appears that the magnification is different for different parts of the lens. It may be mentioned that if we consider the image formation by different zones of a lens, then the spherical aberration arises due to the fact that different zones have different powers and coma arises due to the fact that different zones have different magnifications. In Fig. 5.15 (a) we have shown only those rays which lie in the meridional plane, i.e. that plane containing the optical axis and the object point. To see the shape of the image one has to consider the complete set of rays.* In Fig. 5.15 (b) we have shown a three-dimensional perspective in which we have considered a set of rays which hit the lens at the same distance from the centre. Rays which intersect the lens at diametrically opposite points focus to a

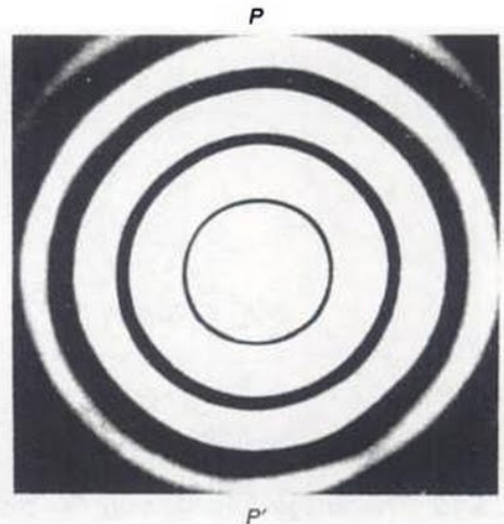
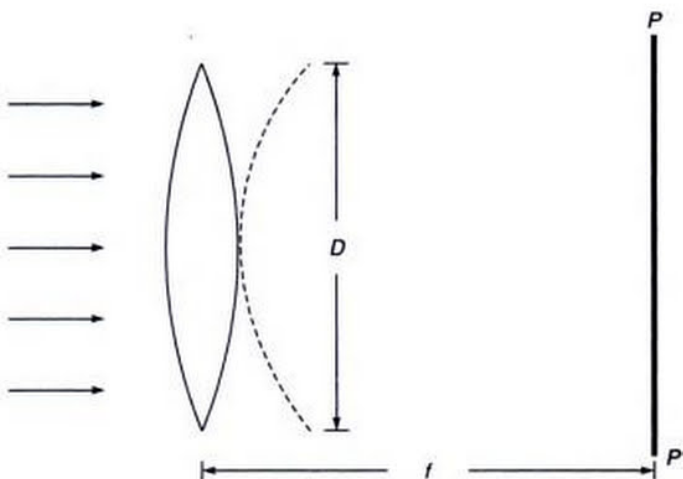


Fig. 5.14 A perfectly spherical wave (converging on the plane PP') will produce an Airy pattern in the image plane.

*It must be mentioned that a proper understanding of the aberrations can only be had by a careful and thorough mathematical analysis. This, however, is beyond the scope of this book; interested readers may look up Refs 1 and 3.

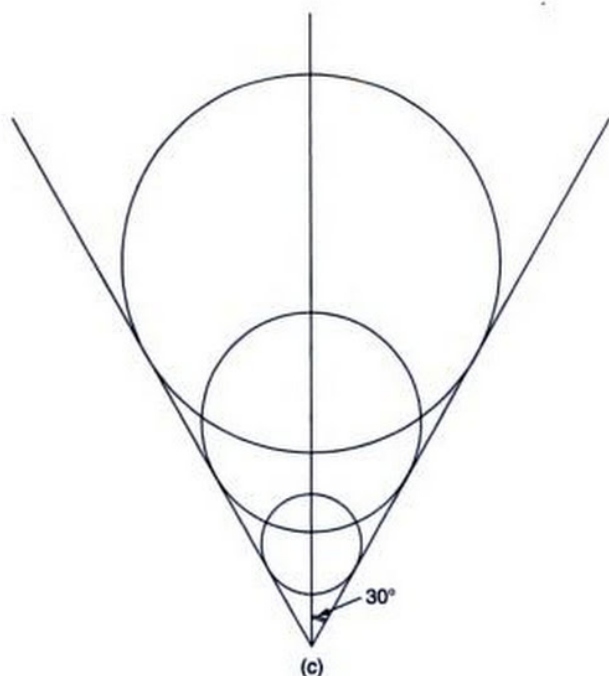
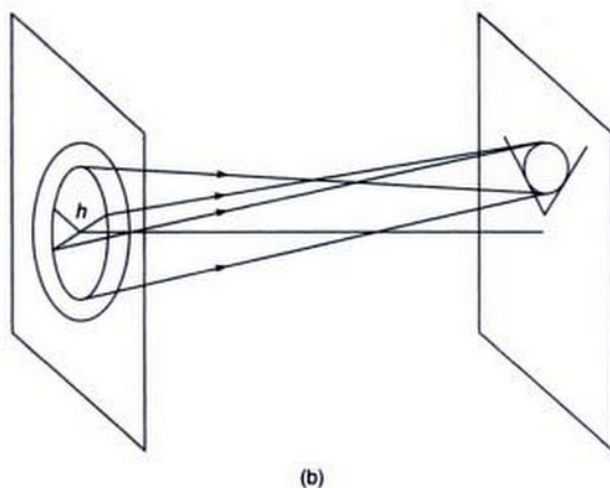
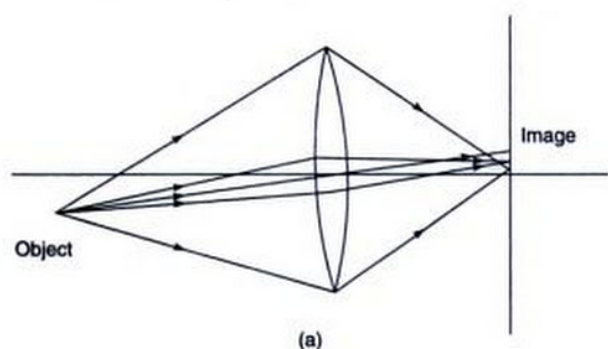


Fig. 5.15 The image formation in the presence of coma. In (a) we have shown only those rays which lie in the meridional plane, (b) a three dimensional perspective is shown. In (c) we have shown the composite image.

single point on the paraxial image plane. These different pairs of rays focus to different points in the image plane such that these foci lie on a circle. The radius of the circle and the distance at which the centre lies from the ideal image point measures the coma. As the radius of the zone [shown as h in Fig. 5.15 (b)] increases, the centre of the circle also shifts away from the ideal image. Thus the composite image will have a form as shown in Fig. 5.15 (c). The image of a point object thus has a comet like appearance and hence the name coma (see Fig. 5.16).



Fig. 5.16 Image of a point source showing coma. (After H.F. Meiners, *Physics Demonstration Experiments*, Vol. II, The Ronald Press Co., New York, 1970; used with permission.)

For a parallel bundle of rays incident on a lens and inclined at an angle θ with the z -axis (see Fig. 5.17), one can show¹ that the coma in the image is given by:

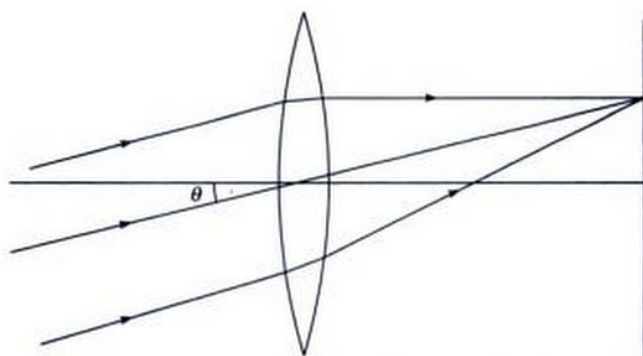


Fig. 5.17 Parallel rays (inclined at an angle θ with the axis) incident on a thin lens.

$$\text{coma} = \frac{3(n-1)}{2} f h^2 \tan^2 \theta \times \left[\frac{(n-1)(2n+1)}{n R_1 R_2} - \frac{n^2 - n - 1}{n^2 R_1^2} - \frac{n}{R_2^2} \right] \quad (39)$$

In Fig. 5.9 we have plotted the variation of coma with the shape factor q . It can immediately be seen that for a lens with $q = +0.8$, coma is zero. It can also be seen that both spherical aberration and coma are close to a minimum for a plano-convex lens (with the convex side facing the incident light) for which $q = 1.0$ and as such plano-convex lenses are extensively used in eye-pieces.

We may mention here that in Sec. 3.11 we had derived the Abbe sine condition which when satisfied, the optical system is free from spherical aberration and coma.

5.3.3 Astigmatism and Curvature of Field

When an optical system is free from spherical aberration and coma then the system will image sharply those object points lying on or near the axis. But for points far away from the axis, the image of a point will not be a point and then the optical system is said to be afflicted with astigmatism.

Consider an object point P far away from the axis. The plane containing the axis and the object point is called the meridional plane and the plane perpendicular to the

meridional plane (containing the axis) is called the **sagittal plane**. Figure 5.18 shows the image formation when the optical system suffers from astigmatism only. The rays in the meridional plane converge at a different point as compared to those in the sagittal plane. For example, rays PA and PB focus at the point T and rays PC and PD focus at a point S which is different from T . Since at the point T , the rays in the sagittal plane have not still focused, one in fact has a focal line which is normal to the meridional plane. This focal line T is called the **tangential focus**. Similarly since at S , the rays in the meridional plane have defocused, one obtains a focal line lying in the tangential plane; this is called the sagittal focal line. The distance between S and T is a measure of astigmatism.

To see the origin of astigmatism one observes that for a point on the axis (when the lens is free from other aberrations) the wavefront emerging from the lens is spherical and thus as the wavefront progresses, it converges to a single point. But when the object point is non-axial, then the emerging wavefront is not spherical and thus as the wavefront converges, it does not focus to a point but to two lines, which are normal to each other and called the tangential and the sagittal focal lines. Somewhere between the two focal lines, the image is circular in shape and is called the circle of least confusion.

The distance between the tangential and the sagittal foci increases as the object point moves away from the axis. Thus the tangential foci and the sagittal foci of points at different distances from the axis lie on two surfaces as

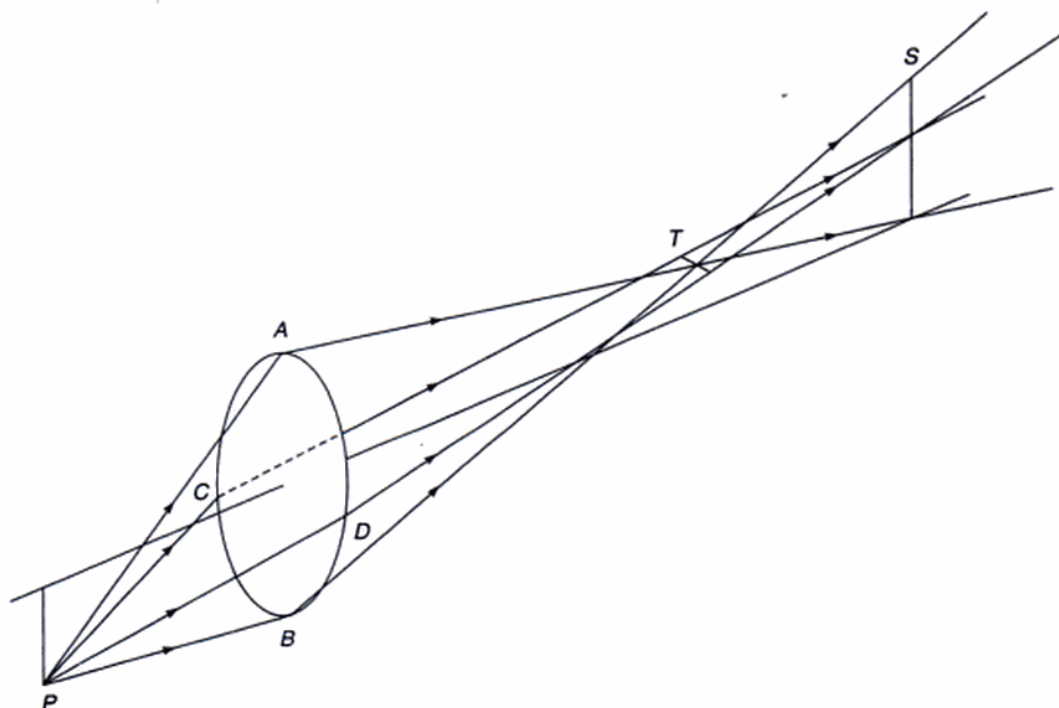


Fig. 5.18 Image formation in the presence of astigmatism.

shown in Fig. 5.19. The optical system will be said to be free from astigmatism when the two surfaces coincide. But even when they coincide it can be shown that the resultant image surface will be curved. This defect of the image is termed **curvature of the field**.

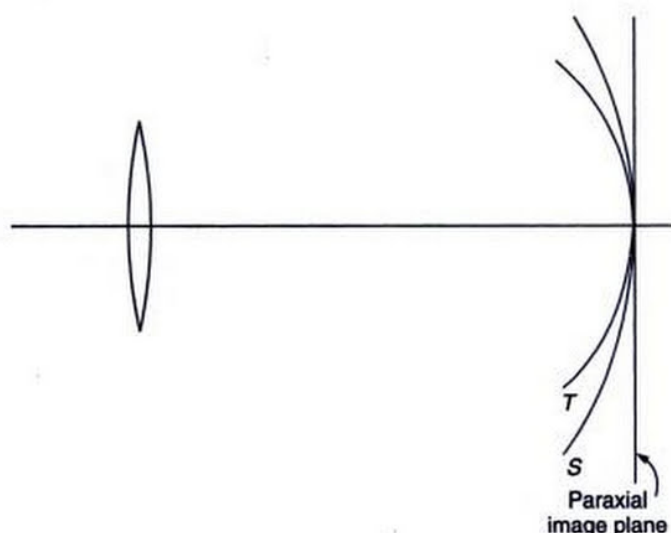


Fig. 5.19 Tangential and sagittal foci.

As an example of image formation in the presence of astigmatism, consider a spoked wheel coaxial with the lens axis as shown in Fig. 5.20 (a). Since on the *T*-surface the image of a point source is a line perpendicular to the meridional plane, on the *T*-surface, the complete rim of the wheel will be in focus while the spokes will be out of focus as shown in Fig. 5.20 (b). Similarly, since on the *S*-surface the image of a point is a line in the meridional plane, the spokes will be in focus and the rim will not be in focus as shown in Fig. 5.20 (c).

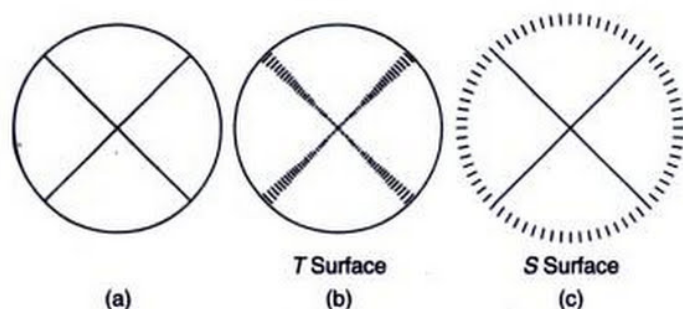


Fig. 5.20 (a) Spoked object coaxial with the axis of the lens; (b) and (c) show images on the *T*-surface and *S*-surface respectively.

5.3.4 Distortion

The last of the Seidel aberrations is called distortion and is caused by non-uniform magnification of the system. When we discussed spherical aberration we had mentioned that for a point object on the axis of the optical system, the images will suffer only from spherical aberration. Similarly, if we have a pinhole on the axis at any plane of the optical system (see Fig. 5.21), then the image will suffer only from distortion. This is because of the fact that corresponding to any point in the object plane, only one of the rays emanating from this point will pass through the pinhole; consequently, all other aberrations will be absent. Obviously, for such a configuration, each point will be imaged as a point but if the system suffers from non-uniform magnification, the image will be distorted. This can be illustrated if we consider the imaging of four equally spaced points *A*, *B*, *C* and *D* which are imaged as *A'*, *B'*, *C'* and *D'* respectively. Mathematical analysis shows that³:

$$X_d = Mx_0 + E(x_0^2 + y_0^2)x_0 \quad (40)$$

and

$$Y_d = My_0 + E(x_0^2 + y_0^2)y_0 \quad (41)$$

where (x_0, y_0) and (X_d, Y_d) represent the coordinates of the object and the image point respectively, *M* represents the magnification of the system and *E* represents the coefficient of distortion. Figure 5.22 (b) corresponds to a negative value of *E* and is known as barrel distortion. The distortion of the image can be easily understood if we consider the imaging of a square grid as shown in Fig. 5.22. Assuming unit magnification (i.e. $M = 1$), the points having coordinates $(0, 0)$, $(h, 0)$, $(2h, 0)$, $(3h, 0)$, $(0, h)$, $(0, 2h)$, $(0, 3h)$, (h, h) , $(h, 2h)$, $(2h, h)$, ... are imaged at $(0, 0)$, $(h + Eh^3, 0)$, $(2h + 8Eh^3, 0)$, $(3h + 27Eh^3, 0)$, $(0, h + Eh^3)$, $(0, 2h + 8Eh^3)$, $(0, 3h + 27Eh^3)$, $(h + Eh^3, h + Eh^3)$, $(h + Eh^3, 2h + 8Eh^3)$, $(2h + 8Eh^3, 2h + 8Eh^3)$, ...

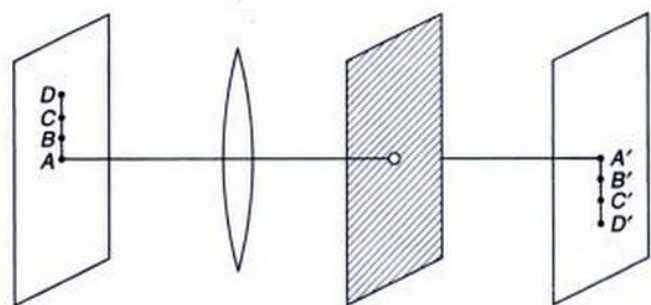


Fig. 5.21 In the presence of a pinhole on the axis, the image suffers only from distortion.

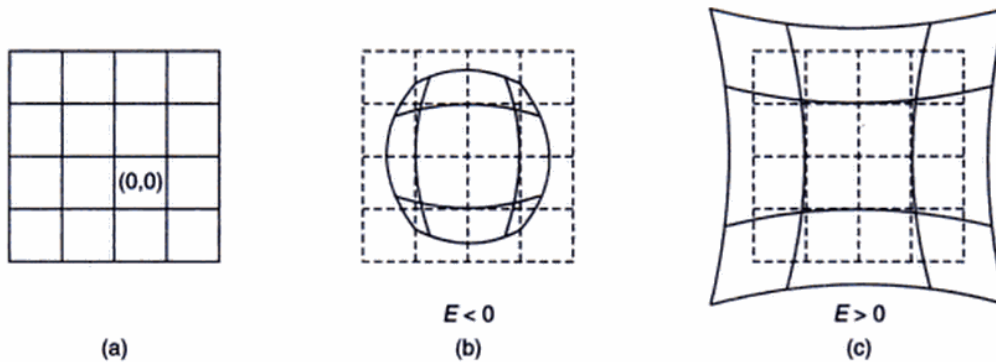


Fig. 5.22 (a) shows the object, (b) represents the image when $E < 0$ and (c) when $E > 0$.

$+8Eh^3, h + Eh^3, \dots$ respectively. If the reader actually plots these points, then for $E < 0$, he would obtain a figure like the one shown in Fig. 5.22 (b). Similarly for $E > 0$, he would obtain Fig. 5.22 (c). Notice that each point is imaged at a point, but the image is distorted because of non-uniform magnification.

SUMMARY

- For a polychromatic source, different wavelength components (after refraction) proceed along different directions and form images at different points; this leads to chromatic aberrations. If we consider two thin lenses made of different materials placed in contact with each other, the focal length of the combination will be the same for blue and red colours if

$$\frac{\omega}{f} + \frac{\omega'}{f'} = 0$$

where

$$\omega = \frac{n_b - n_r}{n - 1} \quad \text{and} \quad \omega' = \frac{n'_b - n'_r}{n' - 1}$$

are known as the dispersive powers. Further,

$$n \equiv \frac{n_b + n_r}{2} \approx n_y, \quad n' \equiv \frac{n'_b + n'_r}{2} \approx n'_y$$

where n_b, n_y and n_r represent the refractive indices for the material of the first lens corresponding to the blue, yellow and red colours respectively. Similarly, n'_b, n'_y and n'_r represent the refractive indices for the second lens. Since ω and ω' are both positive, f and f' must be of opposite signs.

- For a lens, the marginal rays (which are incident near the periphery of the lens) focus at a point which is different than the focal point of paraxial rays. The distance along the axis between the paraxial image point and the image corresponding to marginal rays

(i.e., rays striking the edge of the lens) is termed longitudinal spherical aberration.

- The spherical aberration of a combination of two thin lenses is a minimum when their separation is equal to the difference in their focal lengths.

PROBLEMS

- 5.1** Consider a plane glass slab of thickness d made of a material of refractive index n , placed in air. By simple application of Snell's law obtain an expression for the spherical aberration of the slab. What are other kinds of aberrations that the image will suffer from?

[Ans: Spherical aberration = $-\frac{(n^2 - 1)dh^2}{2n^3u^2}$, where h is

the height at which the ray strikes the slab, and u is the distance of the object point from the front surface of the slab.]

- 5.2** Why can't you obtain an expression for the spherical aberration of a plane glass slab from Eq. (27) by tending R_1, R_2 to ∞ ?
- 5.3** Obtain an expression for the chromatic aberration in the image formed by a plane glass slab.
- 5.4** Does the image formed by a plane mirror suffer from any aberration?
- 5.5** Calculate the longitudinal spherical aberration of a thin plano-convex lens made of a material of refractive index 1.5 and whose curved surface has a radius of curvature of 10 cm, for rays incident at a height of 1 cm. Compare the values of the aberration when the convex side and the plane side face the incident light.
- 5.6** Consider a lens made up of a material of refractive index 1.5 with a focal length 25 cm. Assuming $h = 0.5$ cm and $\theta = 45^\circ$, obtain the spherical aberration.

tion and coma for the lens for various values of the shape factor q and plot the variation in a manner similar to that shown in Fig. 5.9.

- 5.7 An achromatic cemented doublet of focal length 25 cm is to be made from a combination of an equiconvex

flint glass lens ($n_b = 1.50529$, $n_r = 1.49776$) and a crown glass lens ($n_b = 1.66270$, $n_r = 1.64357$). Calculate the radii of curvatures of the different surfaces and the focal lengths of each of the two lenses.

REFERENCES AND SUGGESTED READINGS

1. M. Born, and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
2. M. Cagnet, M. Francon and J.C. Thierr, *Atlas of Optical Phenomena*, Springer-Verlag, Berlin, 1962.
3. A. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978. [Reprinted by Macmillan India, New Delhi.]
4. H.H Hopkins, *Wave Theory of Aberrations*, Oxford University Press, London, 1950.
5. C.J. Smith, 'A Degree Physics', Part III, *Optics*, Edward Arnold Publishers, London, 1960.
6. W.T. Welford, *Geometrical Optics*, North Holland Publishing Co., Amsterdam, 1962.
7. W.T. Welford, *Aberrations of the Symmetrical Optical System*, Academic Press, New York, 1974.

PART 2

Vibrations and Waves

This part (consisting of five chapters) discusses many interesting experiments like the physics behind ionospheric reflection, redness of the setting sun, water waves, shock waves, pulse dispersion, etc. Chapter 6 starts with simple harmonic motion (which is the most fundamental vibration associated with wave motion) and is followed by a derivation of the refractive index variation with frequency. Chapter 7 is essentially mathematics and discusses Fourier series and Fourier transforms which are extensively used in studying the distortion of optical pulses as it propagates through dispersive media (Chapter 8). The derivation and solutions of the wave equation represents the basic physics of wave propagation which have been discussed in Chapter 9. Chapter 10 discusses Huygens' principle which is used to derive the laws of reflection and Snell's law of refraction.

Chapter 6

Simple Harmonic Motion, Forced Vibrations and Origin of Refractive Index

The correct picture of an atom, which is given by the theory of wave mechanics, says that, so far as problems involving light are concerned, the electrons behave as though they were held by springs. So we shall suppose that the electrons have a linear restoring force which, together with their mass m , makes them behave like little oscillators, with a resonant frequency ω_0 The electric field of the light wave polarizes the molecules of the gas, producing oscillating dipole moments. The acceleration of the oscillating charges radiates new waves of the field. This new field, interfering with the old field, produces a changed field which is equivalent to a phase shift of the original wave. Because this phase shift is proportional to the thickness of the material, the effect is equivalent to having a different phase velocity in the material.

— Richard Feynman in Feynman Lectures in Physics, Vol. I

6.1 INTRODUCTION

The most fundamental vibration associated with wave motion is the simple harmonic motion; in Sec. 6.2 we will discuss simple harmonic motion and in Sec. 6.3 we will discuss the effects (on the vibratory motion) due to damping. If a periodic force acts on a vibrating system, the system undergoes what are known as forced vibrations; in Sec. 6.4 we will study such vibrations which will allow us to understand the origin of refractive index (see Sec. 6.5) and even Rayleigh scattering (see Sec. 6.6), which is responsible for the red colour of the setting (or rising) sun and blue colour of the sky.

6.2 SIMPLE HARMONIC MOTION

A periodic motion is a motion which repeats itself after regular intervals of time and the simplest kind of periodic motion is a simple harmonic motion in which the displacement varies sinusoidally with time. In order to understand simple harmonic motion we consider a point P rotating on the circumference of a circle of radius a with an angular velocity ω (see Fig. 6.1). We choose the center of the circle as our origin and we assume that at $t = 0$ the point P lies on the x -axis (i.e., at the point P_0). At an arbitrary

time t the point will be at the position P where $\angle POP_0 = \omega t$.

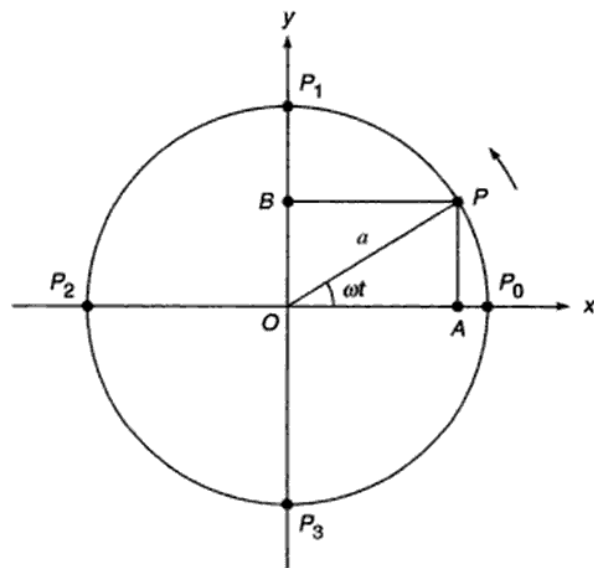


Fig. 6.1 The point P is rotating in the anticlockwise direction on the circumference of a circle of radius a , with uniform angular velocity ω . The foot of the perpendicular on any one of the diameters executes simple harmonic motion. P_0 is the position of the point at $t = 0$.

Let A be the foot of the perpendicular from the point P on the x -axis. Clearly, the distance

$$OA = a \cos \omega t \quad (1)$$

and as the point P rotates on the circumference of the circle, the point A moves to and fro about the origin on the diameter. When the point P is at P_1 , then the foot of the perpendicular is at O . This can also be seen from Eq. (1) because when P coincides with P_1 , $\omega t = \pi/2$ and hence $a \cos \omega t = a \cos \pi/2 = 0$. As the point still moves further, the foot of the perpendicular would lie on the other side of the origin and thus OA would be negative as is also evident from Eq. (1) because ωt would then be greater than $\pi/2$. When P coincides with P_2 , then $OA = OP_2 = -a$. When the point P moves from P_2 to P_3 , OA starts decreasing and it finally goes to zero when P coincides with P_2 . After P crosses P_3 , OA starts increasing again and finally acquires the value a when P coincides with P_0 . After crossing the point P_0 , the motion repeats itself.

A motion in which the displacement varies sinusoidally with time [as in Eq. (1)] is known as a *simple harmonic motion*. Thus, when a point rotates on the circumference of a circle with a uniform angular velocity, the foot of the perpendicular on any one of its diameters will execute simple harmonic motion. The quantity a is called the amplitude of the motion, and the period of the motion, T , will be the time required to complete one revolution. Since the angular velocity is ω , the time taken for one complete revolution will be $2\pi/\omega$. Thus,

$$T = \frac{2\pi}{\omega} \quad (2)$$

The inverse of the time period is known as the frequency:

$$\nu = \frac{1}{T} = \frac{\omega}{2\pi}$$

or

$$\omega = 2\pi\nu \quad (3)$$

It should be pointed out that we could as well have studied the motion of the point B , which is the foot of the perpendicular from the point P on the y -axis. The distance OB is given by (see Fig. 6.1)

$$OB = y = a \sin \omega t \quad (4)$$

We had conveniently chosen $t = 0$ as the time when P was on the x -axis. The choice of the time $t = 0$ is arbitrary and we could have chosen time $t = 0$ to be the instant when P was at P' (see Fig. 6.2). If the angle $\angle P'OX = \theta$ then the projection on the x -axis at any time t would be given by

$$OA = x = a \cos (\omega t + \theta) \quad (5)$$

The quantity $(\omega t + \theta)$ is known as the phase of the motion and θ represents the initial phase. It is obvious from the

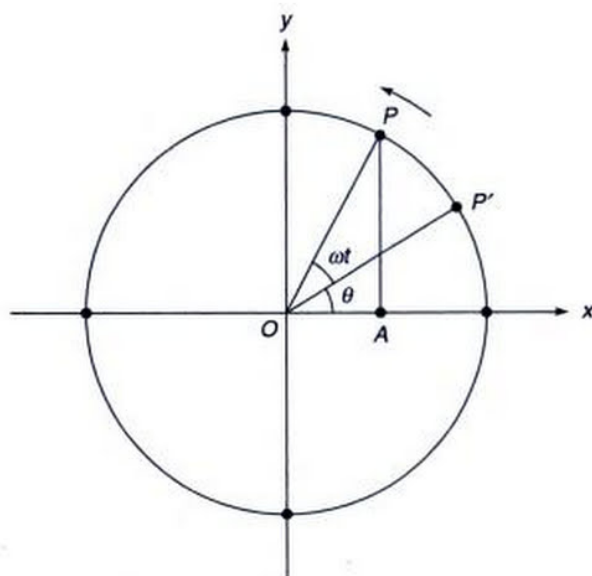


Fig. 6.2 At $t = 0$, the point P is at P' and therefore, the initial phase is θ .

above discussion that the value of θ is quite arbitrary and depends on the instant from which we start measuring time.

We next consider two points P and Q rotating on the circle with the same angular velocity and P' and Q' be their respective positions at $t = 0$. Let the angles $\angle P'OX$ and $\angle Q'OX$ be θ and ϕ respectively (see Fig. 6.3). Clearly at an arbitrary time t , the distance of the foot of the perpendiculars from the origin would be

$$x_P = a \cos (\omega t + \theta) \quad (6a)$$

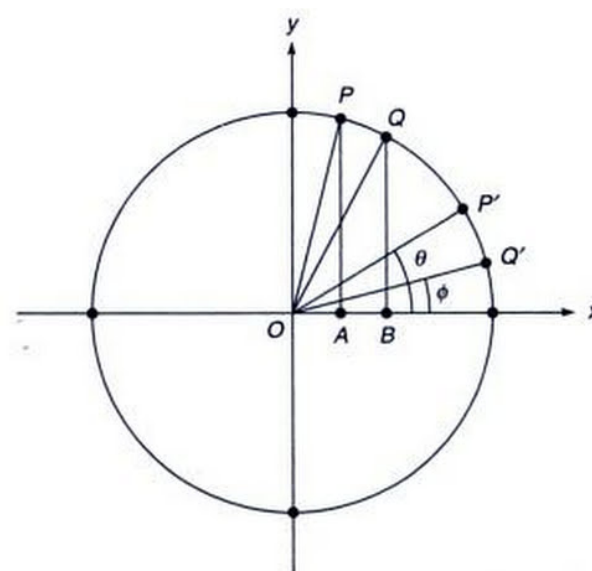


Fig. 6.3 The points A and B execute simple harmonic motions with the same frequency ω . The initial phases of A and B are θ and ϕ respectively.

$$x_Q = a \cos(\omega t + \phi) \quad (6b)$$

The quantity

$$(\omega t + \theta) - (\omega t + \phi) = \theta - \phi \quad (7)$$

represents the phase difference between the two simple harmonic motions and if $\theta - \phi = 0$ (or an even multiple of π) the motions are said to be in phase, and if $\theta - \phi = \pi$ (or an odd multiple of π) the motions are said to be out of phase. If we choose a different origin of time, the quantities θ and ϕ would change by the same additive constant; consequently, the phase difference ($\theta - \phi$) is independent of the choice of the instant $t = 0$.

Thus the displacement of a particle, which executes simple harmonic motion, can be written as:

$$x = a \sin(\omega t + \theta) \quad (8)$$

Therefore, the velocity and the acceleration of the particle would be given by the following equations:

$$v = \frac{dx}{dt} = a\omega \cos(\omega t + \theta) \quad (9)$$

and

$$f = \frac{d^2x}{dt^2} = -a\omega^2 \sin(\omega t + \theta)$$

or,

$$f = \frac{d^2x}{dt^2} = -\omega^2 x \quad (10)$$

Equation (10) shows that the acceleration of the particle is proportional to the displacement and the negative sign indicates that the acceleration is always directed towards the origin. Equation (10) can be used to define the simple harmonic motion as the motion of a particle in a straight line in which the acceleration is proportional to the displacement from a fixed point (on the straight line) and always directed towards the fixed point. (Here the point $x = 0$ is the fixed point and is usually referred to as the equilibrium position.) If we multiply Eq. (10) by the mass of the particle, then we obtain the following expression for the force acting on the particle:

$$F = mf = -m\omega^2 x$$

or

$$F = -kx \quad (11)$$

where $k (= m\omega^2)$ is known as the *force constant*. We could have equally well started from Eq. (11) and obtained simple harmonic motion. This can easily be seen by noting that since the force is acting in the x -direction, the equation of motion would be

$$m \frac{d^2x}{dt^2} = F = -kx$$

or

$$\frac{d^2x}{dt^2} + \frac{k}{m} x = 0$$

or

$$\frac{d^2x}{dt^2} + \omega^2 x = 0 \quad (12)$$

where $\omega^2 = k/m$. The general solution of Eq. (12) can be written in the form

$$x = A \sin \omega t + B \cos \omega t \quad (13)$$

which can be rewritten in either of the following forms:

$$x = a \sin(\omega t + \theta) \quad (14)$$

or

$$x = a \cos(\omega t + \theta) \quad (15)$$

which describes a simple harmonic motion.

6.2.1 Examples of Simple Harmonic Motion

In this section we will discuss three simple examples of simple harmonic motion.

(a) The Simple Pendulum

The simplest example of simple harmonic motion is the motion of the bob of a simple pendulum in the gravitational field. If the bob of the pendulum is displaced slightly, from the equilibrium position (see Fig. 6.4) then the forces acting on the bob are the gravitational force mg acting vertically downwards and the tension T , in the direction $B'A$. In the equilibrium position (AB) the tension is equal and opposite to the gravitational force. However, in the displaced position the tension T is not in the direction of the gravitational force and if we resolve the gravitational force along the direction of the string and perpendicular to it, we see that the component $mg \cos \theta$ balances the tension in the string and the component $mg \sin \theta$ is the restoring force. The motion of the bob is along the arc of a circle but if the length of the pendulum is large and the angle θ is small, the motion can be assumed to be approximately in a straight line [see Fig. 6.4(b)]. Under such an approximation we may assume that this force is always directed towards the point B and the magnitude of this force will be*

$$mg \sin \theta \approx mg \frac{x}{l} \quad (16)$$

*We will be assuming that θ is small so that $\sin \theta \approx \theta$, where θ is in radians. The above approximation is valid for $\theta \leq 0.07$ radians ($\approx 4^\circ$).

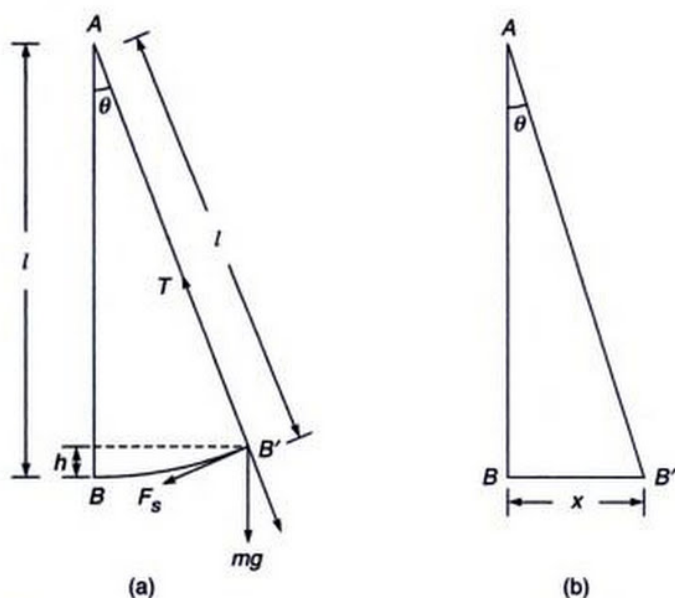


Fig. 6.4 (a) The forces on the bob of the pendulum when it is displaced from its equilibrium position. The restoring force is F_s which is equal to $mg \sin \theta$. (b) If the angle θ is small, the motion of the bob can be approximately assumed to be in a straight line.

Thus the equation of motion will be

$$F = m \frac{d^2 x}{dt^2} = -mg \frac{x}{l} \quad (17)$$

or

$$\frac{d^2 x}{dt^2} + \omega^2 x = 0 \quad (18)$$

where $\omega^2 = g/l$. Equation (18) is of the same form as Eq. (12); thus the motion of the bob is simple harmonic with its time period given by the following equation:

$$T = \frac{2\pi}{\omega} = 2\pi \sqrt{\frac{l}{g}} \quad (19)$$

It should be pointed out that the expression for the time period is fairly accurate (i.e., the motion is approximately simple harmonic) as long as $\theta \leq 4^\circ$.

We next consider the motion of two identical simple pendulums vibrating with the same amplitude a (see Fig. 6.5). Let, at $t = 0$, the bob of one of the pendulums be at its extreme right position, moving towards the right [Fig. 6.5(b)]. If we measure the displacement from the equilibrium positions of the pendulums, then the displacements would be given by

$$\begin{aligned} x_1 &= a \cos \omega t \\ x_2 &= a \sin \omega t = a \cos \left(\omega t - \frac{\pi}{2} \right) \end{aligned} \quad (20)$$

Thus the two bobs execute simple harmonic motion with a phase difference of $\pi/2$ and in fact the first pendulum is ahead in phase by $\pi/2$. It may be mentioned that in Fig. 6.5(b), if the bob was moving towards the left, then the equation of motion would have been

$$x_2 = -a \sin \omega t = a \cos \left(\omega t + \frac{\pi}{2} \right)$$

and then the second pendulum would have been ahead of phase by $\pi/2$. Since, in general, the displacement of the bob of the pendulum can be written as

$$x = a \cos (\omega t + \phi), \quad (21)$$

the velocity of the particle would be given by

$$\frac{dx}{dt} = -a\omega \sin (\omega t + \phi) \quad (22)$$

Thus the kinetic energy of the mass would be

$$\begin{aligned} T &= \frac{1}{2} m \left(\frac{dx}{dt} \right)^2 \\ &= \frac{1}{2} m a^2 \omega^2 \sin^2 (\omega t + \phi) \end{aligned} \quad (23)$$

Comparing Eqs (21) and (23), we see that when the particle is at its extreme positions, the kinetic energy is zero and when the particle passes through equilibrium position, the kinetic energy is maximum. At the extreme positions, the kinetic energy gets transformed into potential energy. From Fig. 6.4(a) it can immediately be seen that

$$\begin{aligned} \text{Potential energy, } V &= mgh = mgl (1 - \cos \theta) \\ &= mgl 2 \sin^2 \frac{\theta}{2} \end{aligned}$$

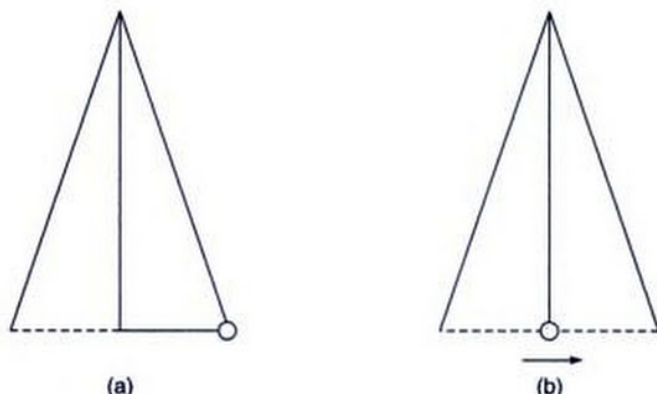


Fig. 6.5 (a) and (b) show the motion of two identical pendulums which are vibrating with the same amplitude but having a phase difference of $\pi/2$. The small circles denote the position of the bobs at $t = 0$.

$$\begin{aligned}
 &\approx 2 mgl \left(\frac{\theta}{2} \right)^2 \\
 &\quad [\theta \text{ measured in radians}] \\
 &\approx \frac{1}{2} mgl \left(\frac{x}{l} \right)^2 = \frac{1}{2} m \left(\frac{g}{l} \right) x^2 \\
 &= \frac{1}{2} m \omega^2 x^2 \quad (24)
 \end{aligned}$$

or

$$V = \frac{1}{2} m \omega^2 a^2 \cos^2(\omega t + \phi) \quad (25)$$

where we have used the fact that $\omega^2 = g/l$. We may mention that the expression for potential energy could have been directly written down by noting the fact that if the potential energies at x and at $x + dx$ are V and $V + dV$, then

$$dV = -F dx = + kx dx \quad (26)$$

Thus

$$V = \int_0^x kx dx = \frac{1}{2} kx^2 \quad (27)$$

where we have assumed the zero of the potential energy to be at $x = 0$. Thus the total energy E would be given by

$$E = T + V = \frac{1}{2} m \omega^2 a^2 \quad (28)$$

which, as expected, is independent of time. We can also see from Eq. (26) that the energy associated with the simple harmonic motion is proportional to the square of the amplitude and the square of the frequency.

(b) Vibrations of a Mass Held by Two Stretched Springs

Another simple example is the motion of a mass m , held by two stretched springs on a smooth table as shown in Fig. 6.6. The two springs are of natural length l_0 [Fig. 6.6(a)] and corresponding to the equilibrium position of the mass, the lengths of the stretched springs are l . If the mass is displaced slightly from the equilibrium position, then the resultant force acting on the mass will be

$$\begin{aligned}
 F &= k[(l - x) - l_0] - k[(l + x) - l_0] \\
 &= -2 kx \quad (29)
 \end{aligned}$$

where k represents the force constant of the spring. Once again we get a force which is proportional to the displacement and directed towards the equilibrium position and consequently, the motion of the mass on the frictionless table will be simple harmonic in nature.

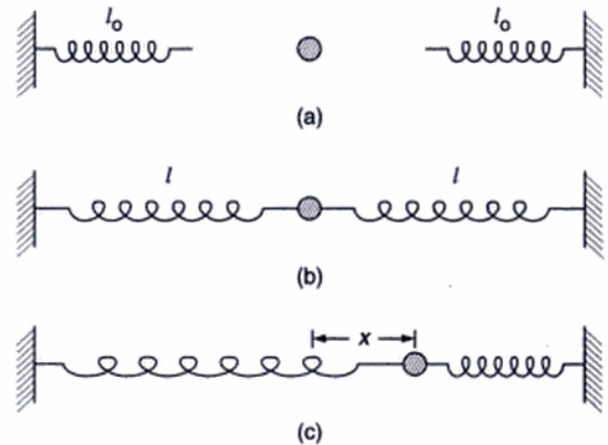


Fig. 6.6 Two springs of natural length l_0 [see (a)] are stretched to a length l [see (b)] to hold the mass. If the mass is displaced by a small distance x from its equilibrium position [see (c)], the mass will execute simple harmonic motion.

(c) Vibrations of a Stretched String

When a stretched string (as in a sonometer) is made to vibrate in its fundamental mode (see Fig. 6.7), then each point on the string executes simple harmonic motion with different amplitudes but having the same initial phase. The displacement can be written in the form

$$y = a \sin \left(\frac{\pi}{L} x \right) \cos \omega t \quad (30)$$

The amplitude is therefore zero at $x = 0$ and at $x = L$ and is maximum at $x = L/2$. On the other hand, if the string is vibrating in its first harmonic, then each point on the first half of the string vibrates out of phase with each point on the other half.

6.3 DAMPED SIMPLE HARMONIC MOTION

In Sec. 6.2, we had shown that for a particle executing SHM, the equation of motion will be of the form



Fig. 6.7 When a string clamped at both the ends is made to vibrate in its fundamental mode, all particles execute simple harmonic motions with same frequency and same initial phase but having different amplitudes.

$$\frac{d^2x}{dt^2} + \omega_0^2 x(t) = 0 \quad (31)$$

the solution of which is given by

$$x(t) = A \cos(\omega_0 t + \theta) \quad (32)$$

where A represents amplitude and ω_0 the angular frequency of motion. Equation (32) tells us that the motion will continue forever. However, we know that in actual practice the amplitude of any vibrating system (like that of a tuning fork) keeps on decreasing and eventually the system stops vibrating. Similarly, the bob of a pendulum comes to rest after a certain period of time. This phenomenon is due to the presence of damping forces which come into play when the particle is in motion. For a vibrating pendulum, the damping forces are primarily due to the viscosity of the surrounding medium. Consequently, the damping forces will be much larger in liquids than in gases. In general, the exact dependence of the damping force on the velocity of the particle is quite complicated; however, as a first approximation we may assume it to be proportional to the velocity of the particle. This is also consistent with the fact that there are no damping forces acting on the particle when it is at rest. In this model, the equation of motion will be given by

$$m \frac{d^2x}{dt^2} = -\Gamma \frac{dx}{dt} - k_0 x \quad (33)$$

where the constant Γ determines the strength of the damping force; the force constant is now denoted by k_0 to avoid confusion with the wave vector k . Equation (33) can be rewritten in the form

$$\frac{d^2x}{dt^2} + 2K \frac{dx}{dt} + \omega_0^2 x(t) = 0 \quad (34)$$

where

$$2K = \frac{\Gamma}{m} \quad \text{and} \quad \omega_0 = \sqrt{\frac{k_0}{m}} \quad (35)$$

In order to solve Eq. (34) we introduce a new variable $\xi(t)$ which is defined by the following equation:

$$x(t) = \xi(t) e^{-Kt} \quad (36)$$

Thus,

$$\frac{dx}{dt} = \left[\frac{d\xi}{dt} - K\xi(t) \right] e^{-Kt}$$

and

$$\frac{d^2x}{dt^2} = \left[\frac{d^2\xi}{dt^2} - 2K \frac{d\xi}{dt} + K^2 \xi(t) \right] e^{-Kt}$$

On substitution in Eq. (34) we get

$$\frac{d^2\xi}{dt^2} + (\omega_0^2 - K^2) \xi(t) = 0 \quad (37)$$

Equation (37) is similar to Eq. (31); however, depending on the strength of the damping force, the quantity $(\omega_0^2 - K^2)$ can be positive, negative or zero. Consequently, we must consider three cases.

Case I ($\omega_0^2 > K^2$)

If the damping is small, ω_0^2 is greater than K^2 , and the solution of Eq. (37) would be of the form

$$\xi(t) = A \cos \left[\sqrt{\omega_0^2 - K^2} t + \theta \right] \quad (38)$$

or

$$x(t) = A e^{-Kt} \cos \left[\sqrt{\omega_0^2 - K^2} t + \theta \right] \quad (39)$$

where A and θ are constants which are determined from the amplitude and phase of the motion at $t = 0$. Equation (39) represents a damped simple harmonic motion (see Fig. 6.8). Notice that the amplitude decreases exponentially with time and the time period of vibration $\left(= 2\pi / \sqrt{\omega_0^2 - K^2} \right)$ is greater than in the absence of damping.

Case II ($K^2 > \omega_0^2$)

If the damping is too large, K^2 is greater than ω_0^2 , and Eq. (37) should be written in the form

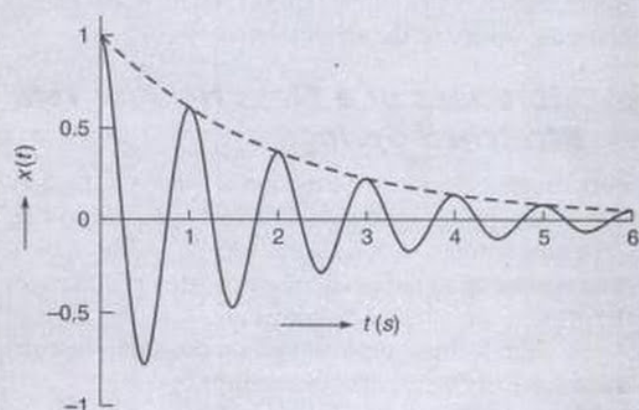


Fig. 6.8 The exponential decrease of amplitude in a damped simple harmonic motion. The figure corresponds to $\frac{2\pi}{\sqrt{\omega_0^2 - K^2}} = 1\text{ s}$ and $K = 0.5\text{ s}^{-1}$.

$$\frac{d^2\xi}{dt^2} - (K^2 - \omega_0^2) \xi(t) = 0 \quad (40)$$

the solution of which is given by

$$\xi(t) = A \exp\left[\sqrt{K^2 - \omega_0^2} t\right] + B \exp\left[-\sqrt{K^2 - \omega_0^2} t\right] \quad (41)$$

Thus,

$$x(t) = A \exp\left[\left(-K + \sqrt{K^2 - \omega_0^2}\right)t\right] + B \exp\left[\left(-K - \sqrt{K^2 - \omega_0^2}\right)t\right] \quad (42)$$

and we can have two kinds of motion; one in which the displacement decreases uniformly to zero, or the other, in which the displacement first increases, reaches a maximum and then decreases to zero (see Fig. 6.9). In either case there are no oscillations and the motion is said to be *overdamped* or *dead beat*. A typical example is the motion of a simple pendulum in a highly viscous liquid (like glycerine) where the pendulum can hardly complete a fraction of the vibration before coming to rest.

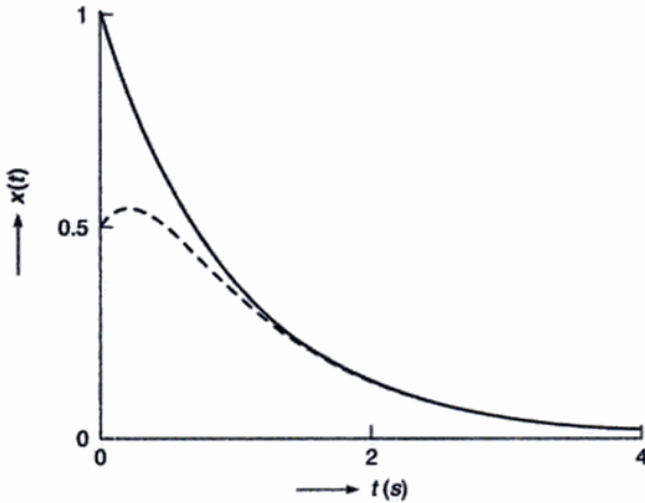


Fig. 6.9 The variation of displacement with time in an overdamped motion. The solid and the dashed curves correspond to $B = 0$ and $B = -A/2$ respectively [see Eq. (42)]. In carrying out the calculations we have assumed $K = 2 \text{ s}^{-1}$ and $\sqrt{K^2 - \omega_0^2} = 1 \text{ s}^{-1}$.

Case III ($K^2 = \omega_0^2$)

When, $K^2 = \omega_0^2$, Eq. (37) becomes

$$\frac{d^2\xi}{dt^2} = 0 \quad (43)$$

the solution of which is given by

$$\xi = At + B \quad (44)$$

Thus

$$x(t) = (At + B) e^{-Kt} \quad (45)$$

The motion is again non-oscillatory and is said to correspond to *critical damping*.

6.4 FORCED VIBRATIONS

We consider the effect of a periodic sinusoidal force (see also Sec. 7.3) on the motion of a vibrating system. If the frequency of the external force is ω then the equation of motion would be [cf. Eq. (33)]:

$$m \frac{d^2x}{dt^2} = F \cos \omega t - \Gamma \frac{dx}{dt} - k_0 x \quad (46)$$

where the first term on the RHS represents the external force; the other terms are the same as in Eq. (33). Equation (46) is rewritten in the form*

$$\frac{d^2x}{dt^2} + 2K \frac{dx}{dt} + \omega_0^2 x(t) = G \cos \omega t \quad (47)$$

where $G = F/m$ and other symbols have been defined in Sec. 6.3. For the particular solution of Eq. (47) we try

$$x(t) = a \cos(\omega t - \phi) \quad (48)$$

Thus,

$$\frac{dx}{dt} = -a\omega \sin(\omega t - \phi)$$

and

$$\frac{d^2x}{dt^2} = -a\omega^2 \cos(\omega t - \phi)$$

Substituting the above forms for $x(t)$, dx/dt and d^2x/dt^2 in Eq. (47), we obtain

$$-a\omega^2 \cos(\omega t - \phi) - 2K a\omega \sin(\omega t - \phi) + a\omega_0^2 \cos(\omega t - \phi) = G \cos[(\omega t - \phi) + \phi] \quad (49)$$

where we have written $G \cos \omega t$ as $G \cos[(\omega t - \phi) + \phi]$.

*Notice that the RHS of Eq. (47) is independent of x ; such an equation is said to be an inhomogeneous equation. An equation of the type given by Eq. (34) is said to be homogeneous.

Thus,

$$a(\omega_0^2 - \omega^2) \cos(\omega t - \phi) - 2Ka\omega \sin(\omega t - \phi) = G \cos(\omega t - \phi) \cos \phi - G \sin(\omega t - \phi) \sin \phi \quad (50)$$

For Eq. (50) to be valid for all values of time we must have

$$a(\omega_0^2 - \omega^2) = G \cos \phi \quad (51)$$

$$2Ka\omega = G \sin \phi \quad (52)$$

If we square and add, we get

$$a = \frac{G}{[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2]^{1/2}} \quad (53)$$

Further

$$\tan \phi = \frac{2K\omega}{(\omega_0^2 - \omega^2)} \quad (54)$$

Since K , ω and a are positive, ϕ is uniquely determined by noting that $\sin \phi$ should be positive, i.e., ϕ must be either in the first or in the second quadrant.

To the solution given by Eq. (48), we must add the solution of the homogeneous equation, Eq. (34). Thus, assuming ω_0^2 to be greater than K^2 (i.e., weak damping), the general solution of Eq. (47) will be of the form

$$x(t) = Ae^{-Kt} \cos[\sqrt{\omega_0^2 - K^2} t - \theta] + a \cos(\omega t - \phi) \quad (55)$$

The first term on the RHS represents the transient solution (corresponding to the natural vibrations of the system) which eventually die out. The second term represents the steady state solution which corresponds to the forced vibrations imposed by the external force. Notice that the frequency of the forced vibrations is the same as that of the external force.

6.4.1 Resonance

The amplitude of the forced vibration,

$$a = \frac{G}{[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2]^{1/2}} \quad (56)$$

depends on the frequency of the driving force and is a maximum when $(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2$ is a minimum, i.e., when

$$\frac{d}{d\omega} [(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2] = 0$$

*There is no resonance condition when $K^2 \geq \frac{1}{2} \omega_0^2$.

or

$$2(\omega_0^2 - \omega^2)(-2\omega) + 8K^2\omega = 0$$

or

$$\omega = \omega_0 \left[1 - \frac{2K^2}{\omega_0^2} \right]^{1/2} \quad (57)$$

Thus the amplitude is maximum* when ω is given by Eq. (57). This is known as *amplitude resonance*. When damping is extremely small, the resonance occurs at a frequency very close to the natural frequency of the system. The variation of the amplitude with ω is shown in Fig. 6.10. Notice that as the damping decreases, the maximum becomes very sharp and the amplitude falls off rapidly as we go away from the resonance. The maximum value of a is given by

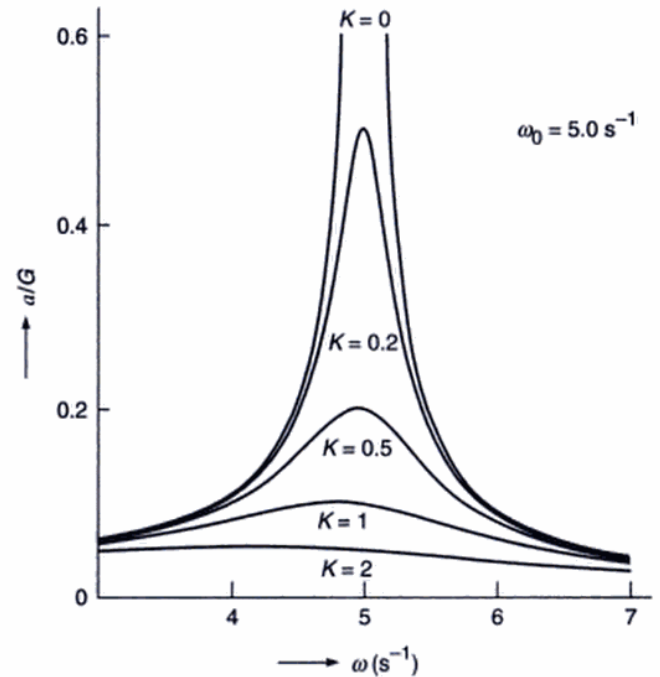


Fig. 6.10 The variation of amplitude with the frequency of the external driving force for various values of K . The calculations correspond to $\omega_0 = 5 \text{ s}^{-1}$ and the values of K are in sec^{-1} . Notice that with increase in damping, the resonance occurs at a smaller value of ω .

$$a_{\max} = \frac{G}{\left[(2K^2)^2 + 4K^2\omega_0^2 \left(1 - \frac{2K^2}{\omega_0^2} \right) \right]^{1/2}}$$

$$= \frac{G}{2K[\omega_0^2 - K^2]^{1/2}} = \frac{G}{2K[\omega^2 + K^2]^{1/2}} \quad (58)$$

Thus, with increase in damping, the maximum occurs at lower values of ω and the resonance becomes less sharper.

In order to discuss the phase of the forced vibrations, we refer to Eq. (54) from where we find that for small damping the phase angle is small unless it is near resonance. For $\omega = \omega_0$, $\tan \phi = \infty$ and $\phi = \pi/2$; i.e., the phase of forced vibrations is $\pi/2$ ahead of the phase of the driving force. As the frequency of the driving force is increased beyond ω_0 , the phase also increases and approaches π (see Fig. 6.11).

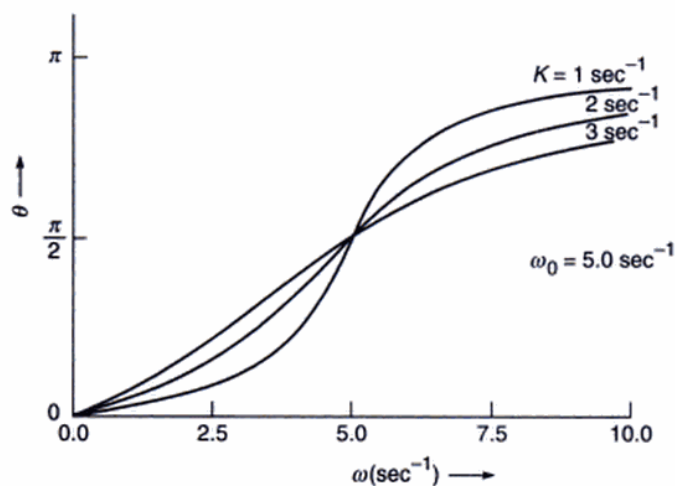


Fig. 6.11 The dependence of the phase of the forced vibration on the frequency of the driving force.

All the salient features of forced vibrations can be easily demonstrated by means of an arrangement shown in Fig. 6.12. In the figure, AC is a metal rod with a movable bob B and LM is a simple pendulum with a bob at M . The metal rod and the simple pendulum are suspended from a string PQ as shown in Fig. 6.12. With B at the bottom, when the rod AC is set in motion, the pendulum LM also vibrates. As the bob B is moved upwards, the time period decreases and the frequency of the rod becomes closer to the natural frequency of the simple pendulum and eventually the resonance condition is satisfied. At resonance, the amplitude of vibration of the simple pendulum is maximum and the phase difference between the vibrations is nearly $\pi/2$; i.e., when the metal rod is at its lowest position and moving towards right, the simple pendulum is at the extreme left position. If

the bob B is further moved upward, the frequency increases, and the amplitude of the forced vibrations decreases.

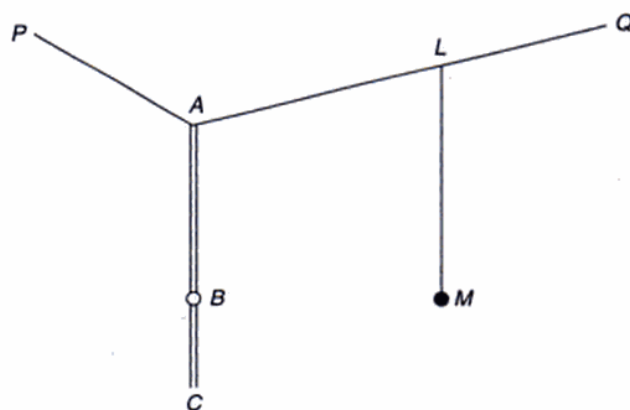


Fig. 6.12 An arrangement for demonstration of forced vibrations.

6.5 ORIGIN OF REFRACTIVE INDEX

In this section we will study the origin of refractive index. We know that an atom consists of a heavy positively charged nucleus surrounded by electrons. In the simplest model of the atom, the electrons are assumed to be bound elastically to their rest positions; thus, when these electrons are displaced by an electric field, a restoring force (proportional to the displacement) will act on the electrons which will tend to return the electrons to their rest positions. In this model, the equation of motion for the electron, in the presence of an external electric field \mathbf{E} , would be

$$m \frac{d^2 \mathbf{x}}{dt^2} + k_0 \mathbf{x} = -q\mathbf{E} \quad (59)$$

or

$$\frac{d^2 \mathbf{x}}{dt^2} + \omega_0^2 \mathbf{x} = -\frac{q}{m} \mathbf{E} \quad (60)$$

where \mathbf{x} represents the position of the electron, m and $-q$ represent the mass and charge of the electron ($q \approx +1.6 \times 10^{-19}$ C), k_0 the force constant and $\omega_0 = (\sqrt{k_0/m})$ represents the frequency of the oscillator. We assume

$$\mathbf{E} = \hat{\mathbf{x}} E_0 \cos(kz - \omega t) \quad (61)$$

i.e., the field is in the x -direction having an amplitude E_0 and propagating in the $+z$ direction; $\hat{\mathbf{x}}$ represents the unit vector in the x -direction and $k = 2\pi/\lambda$, λ representing the wavelength. Thus

$$\frac{d^2 x}{dt^2} + \omega_0^2 x = -\frac{qE_0}{m} \cos(kz - \omega t) \quad (62)$$

where we have replaced the vectors by the corresponding scalar quantities because the displacement and the electric field are in the same direction. Except for the damping term, Eq. (62) is similar to Eq. (46) and therefore, the solution corresponding to the forced vibrations will be given by*

$$x = -\frac{qE_0}{m(\omega_0^2 - \omega^2)} \cos(kz - \omega t) \quad (63)$$

In the simplest model of the atom, the centre of the negative charge (due to the electrons) is assumed to be at the centre of the nucleus. In the presence of an electric field, the centre of the negative charge gets displaced from the nucleus which results in a finite value of the dipole moment of the atom. In particular, if we have a positive charge $+q$ at the origin and a negative charge $-q$ at a distance x , then the dipole moment would be $-qx$; thus, if there are N dispersion-electrons** per unit volume then the polarization (i.e., dipole moment per unit volume) would be given by

$$\begin{aligned} \mathbf{P} &= -Nq \mathbf{x} = \frac{Nq^2}{m(\omega_0^2 - \omega^2)} \mathbf{E} \\ &= \chi \mathbf{E} \end{aligned} \quad (64)$$

where

$$\chi = \frac{Nq^2}{m(\omega_0^2 - \omega^2)} \quad (65)$$

is known as the electric susceptibility of the material. The dielectric permittivity is therefore given by (see Chapter 20)

$$\epsilon = \epsilon_0 + \chi \quad (66)$$

or

$$\frac{\epsilon}{\epsilon_0} = 1 + \frac{Nq^2}{m\epsilon_0(\omega_0^2 - \omega^2)} \quad (67)$$

Now, ϵ/ϵ_0 is the dielectric constant, which is equal to the square of the refractive index (see Chapter 20). Thus

$$n^2 = 1 + \frac{Nq^2}{m\epsilon_0\omega_0^2} \left[1 - \frac{\omega^2}{\omega_0^2} \right]^{-1} \quad (68)$$

showing that the refractive index depends on the frequency; this is known as dispersion. Assuming that the characteristic frequency ω_0 lies in the far ultraviolet [see Eq. (74)]***, the

quantity $\left[1 - \frac{\omega^2}{\omega_0^2} \right]^{-1}$ is positive in the entire visible region.

Further, as ω increases, n^2 also increases, i.e., the refractive index increases with frequency; this is known as normal dispersion. If we further assume $\omega/\omega_0 \ll 1$, then

$$\left[1 - \frac{\omega^2}{\omega_0^2} \right]^{-1} \approx 1 + \frac{\omega^2}{\omega_0^2}$$

and

$$\begin{aligned} n^2 &= 1 + \frac{Nq^2}{m\epsilon_0\omega_0^2} \left[1 + \frac{\omega^2}{\omega_0^2} \right] \\ &\approx 1 + \frac{Nq^2}{m\epsilon_0\omega_0^2} + \frac{4\pi^2 c^2 Nq^2}{m\epsilon_0\omega_0^4} \frac{1}{\lambda_0^2} \end{aligned} \quad (69)$$

where $\lambda_0 = 2\pi c/\omega_0$ is the free space wavelength. Equation (69) can be written in the form

$$n^2 = A + \frac{B}{\lambda_0^2} \quad (70)$$

which is the well-known 'Cauchy relation'. For hydrogen, the experimental variation of n^2 with λ_0 is approximately given by the following relation:

$$n^2 = 1 + 2.721 \times 10^{-4} + \frac{2.11 \times 10^{-18}}{\lambda_0^2} \quad (71)$$

where the wavelength is measured in metres; the above numbers correspond to 0°C and 76 cm of Hg^9 . Thus,

$$\frac{Nq^2}{m\epsilon_0\omega_0^2} = 2.721 \times 10^{-4} \quad (72)$$

and

$$\frac{4\pi^2 c^2 Nq^2}{m\epsilon_0\omega_0^4} = 2.11 \times 10^{-18} \text{ m}^2 \quad (73)$$

If we divide the second equation by the first, we would get

$$\frac{4\pi^2 c^2}{\omega_0^2} = \frac{2.11 \times 10^{-18}}{2.721 \times 10^{-4}}$$

$$\text{or} \quad \nu_0 = \frac{\omega_0}{2\pi} \approx 3 \times 10^{15} \text{ s}^{-1} \quad (74)$$

*Notice that in the absence of damping (i.e., when $\Gamma = 0$), $\phi = 0$; see Eq. (54).

**The number of 'dispersion-electrons' in a molecule of an ideal gas is the valence number of the molecules. This number is 2 for H_2 , 6 for N_2 , etc.

***This also follows from the fact that according to classical electrodynamics, an oscillating dipole vibrating with frequency ω_0 will radiate electromagnetic waves with frequency ω_0 ; and as an example if we consider hydrogen, then $\hbar \omega_0 = 13.6 \text{ eV}$ from which one obtains $\omega_0 = 2 \times 10^{16} \text{ s}^{-1}$. This frequency corresponds to the far ultraviolet.

which is indeed in the ultraviolet region. One can eliminate ω_0 from Eqs (72) and (73) to obtain

$$\frac{Nq^2}{4\pi^2 c^2 \epsilon_0 m} = 3 \times 10^{10} \text{ m}^{-2} \quad (75)$$

Now at NTP, 22400 cc of H_2 contains 6×10^{23} molecules; thus,

$$N = 2 \times \frac{6 \times 10^{23}}{22400 \times 10^{-5}} \text{ m}^{-3} \approx 5 \times 10^{25} \text{ m}^{-3}$$

where the factor 2 arises from the fact that a hydrogen molecule consists of two electrons. Hence,

$$\begin{aligned} \frac{Nq^2}{4\pi^2 c^2 \epsilon_0 m} &= \frac{5 \times 10^{25} \times (1.6 \times 10^{-19})^2}{4 \times \pi^2 \times 9 \times 10^{16} \times 8.85 \times 10^{-12} \times 9.1 \times 10^{-31}} \\ &\approx 4 \times 10^{10} \text{ m}^{-2} \end{aligned}$$

which qualitatively agrees with Eq. (75).

It is of interest to mention that for a gas of free electrons (as we have in the upper atmosphere) there is no restoring force and we must set $\omega_0 = 0$. Thus the expression for the refractive index becomes [see Eq. (67)]

$$n^2 = 1 - \frac{Nq^2}{m\epsilon_0 \omega^2} \quad (76)$$

where N represents the density of free electrons. Equation (75) shows that the refractive index is less than unity; however, this does not imply that one can send signals faster than the speed of light in free space (see Chapter 8). To quote Feynman:

For free electrons, $\omega_0 = 0$ (there is no elastic restoring force). Setting $\omega_0 = 0$ in our dispersion equation yields the correct formula for the index of refraction for radiowaves in the stratosphere, where N is now to represent the density of free electrons (number per unit volume) in the stratosphere. But let us look again at the equation, if we beam X-rays on the matter, or radiowaves (or any electric waves) on free electrons, the term $(\omega_0^2 - \omega^2)$ become negative, and we obtain the result that n is less than one. That means that the effective speed of the waves in the substance is faster than c ! Can that be correct? It is correct. In spite of the fact that it is said that you cannot send signals any faster than the speed of light, it is nevertheless true that the index of refraction of materials at a particular frequency can be either greater or less than 1.

Equation (76) is usually written in the form

$$n^2 = 1 - \left(\frac{\omega_p}{\omega} \right)^2 \quad (77)$$

where

$$\omega_p = \left(\frac{Nq^2}{m\epsilon_0} \right)^{1/2} \quad (78)$$

is known as the plasma frequency. Notice that for $\omega < \omega_p$, the refractive index is purely imaginary which gives rise to attenuation and for $\omega > \omega_p$, the refractive index is real. Indeed in 1933, Wood discovered that alkali metals are transparent to ultraviolet light. For example, for sodium if we assume that the refractive index is primarily due to the free electrons and that there is one free electron per atom then

$$N = \frac{6 \times 10^{23} \times 0.9712}{22.99} \approx 2.535 \times 10^{22} \text{ cm}^{-3}$$

where we have assumed that the atomic weight of Na is 22.99 and its density is 0.9712 g/cm^3 . Substituting the values of $m \approx 9.109 \times 10^{-31} \text{ kg}$, $q \approx 1.602 \times 10^{-19} \text{ C}$ and $\epsilon_0 \approx 8.854 \times 10^{-12} \text{ C/N-m}^2$ we would get

$$\lambda_p \left(= \frac{2\pi c}{\omega_p} \right) \approx 2098 \text{ \AA}$$

Thus for $\lambda < 2098 \text{ \AA}$, the refractive index of Na becomes real and the metal would become transparent; the corresponding experimental value is 2100 \AA . The theoretical and experimental values of λ_p for Li, K and Rb are discussed in Problem 6.7.

As mentioned above, Eq. (76) gives the correct dependence of the refractive index of the stratosphere for radio-waves; in Sec. 2.4.3 we had used Eq. (76) to study reflection of electromagnetic waves by the ionosphere.

Returning to Eq. (68), we note that as $\omega \rightarrow \omega_0$, the refractive index tends to ∞ . This is due to the fact that we have neglected the presence of damping forces in our treatment. If we do take into account the damping forces, Eq. (62) would modify to [see Eq. (46)]

$$m \frac{d^2 x}{dt^2} + \Gamma \frac{dx}{dt} + k_0 x = qE_0 \cos(kz - \omega t) \quad (79)$$

In order to derive an expression for the refractive index, it is more convenient to rewrite the above equation in the form

$$\frac{d^2 x}{dt^2} + 2K \frac{dx}{dt} + \omega_0^2 x = \frac{qE_0}{m} e^{i(kz - \omega t)} \quad (80)$$

where the solution of Eq. (79) will be the real part of the solution of Eq. (80). The solution of the homogeneous equation will give the transient behaviour which will die

out as $t \rightarrow \infty$ (see Sec. 6.4); the steady state solution will correspond to frequency ω . Thus, if we substitute a solution of the type

$$x(t) = A e^{i(kz - \omega t)} \quad (81)$$

in Eq. (80), we would obtain

$$(-\omega^2 - 2iK\omega + \omega_0^2)A = \frac{qE_0}{m}$$

or*

$$A = \frac{qE_0}{m[\omega_0^2 - \omega^2 - 2iK\omega]} \quad (82)$$

Thus we get

$$\mathbf{P} = \frac{Nq^2}{m[\omega_0^2 - \omega^2 - 2iK\omega]} \mathbf{E} \quad (83)$$

The electric susceptibility would therefore be given by

$$\chi = \frac{Nq^2}{m[\omega_0^2 - \omega^2 - 2iK\omega]}$$

Thus

$$\begin{aligned} n^2 &= \frac{\epsilon}{\epsilon_0} = 1 + \frac{\chi}{\epsilon_0} \\ &= 1 + \frac{Nq^2}{m\epsilon_0[\omega_0^2 - \omega^2 - 2iK\omega]} \end{aligned} \quad (84)$$

Notice that the refractive index is complex, which implies absorption of the propagating electromagnetic wave. Indeed, if we write

$$n = \eta + i\kappa \quad (85)$$

where η and κ are real numbers, then the wave number k , which equals $n\omega/c$, would be given by

$$k = (\eta + i\kappa) \frac{\omega}{c} \quad (86)$$

If we consider a plane electromagnetic wave propagating in the $+z$ direction, then its z and t dependence would be of the form $\exp[i(kz - \omega t)]$; consequently

$$\begin{aligned} \mathbf{E} &= \mathbf{E}_0 e^{i(kz - \omega t)} \\ &= \mathbf{E}_0 \exp\left[i\left\{(\eta + i\kappa) \frac{\omega}{c} z - \omega t\right\}\right] \\ &= \mathbf{E}_0 \exp\left[-i\omega\left(t - \frac{\eta z}{c}\right) - \frac{\kappa\omega}{c} z\right] \end{aligned} \quad (87)$$

which shows an exponential attenuation of the amplitude. This should not be unexpected because damping causes a loss of energy.

In order to obtain expressions for η and κ , we substitute the expression for n from Eq. (85) in Eq. (84) to obtain

$$\begin{aligned} (\eta + i\kappa)^2 &= 1 + \frac{Nq^2(\omega_0^2 - \omega^2 + 2iK\omega)}{m\epsilon_0(\omega_0^2 - \omega^2 - 2iK\omega)(\omega_0^2 - \omega^2 + 2iK\omega)} \\ &= 1 + \frac{Nq^2(\omega_0^2 - \omega^2 + 2iK\omega)}{m\epsilon_0(\omega_0^2 - \omega^2 - 2iK\omega)(\omega_0^2 - \omega^2 + 2iK\omega)} \end{aligned}$$

or

$$\eta^2 - \kappa^2 = 1 + \frac{Nq^2(\omega_0^2 - \omega^2)}{m\epsilon_0[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2]} \quad (88)$$

and

$$2\eta\kappa = \frac{Nq^2}{m\epsilon_0} \frac{2K\omega}{[(\omega_0^2 - \omega^2)^2 + 4K^2\omega^2]} \quad (89)$$

The above equations can be rewritten in the form

$$\eta^2 - \kappa^2 = 1 - \frac{\alpha\Omega}{[\Omega^2 + \beta^2(\Omega + 1)]} \quad (90)$$

and

$$2\eta\kappa = \frac{\alpha\beta\sqrt{1+\Omega}}{[\Omega^2 + \beta^2(\Omega + 1)]} \quad (91)$$

where we have introduced the following dimensionless parameters:

$$\alpha = \frac{Nq^2}{m\epsilon_0\omega_0^2}; \quad \Omega = \frac{\omega^2 - \omega_0^2}{\omega_0^2} \quad \text{and} \quad \beta = \frac{2K}{\omega_0}$$

The qualitative variations of $\eta^2 - \kappa^2$ and $2\eta\kappa$ with Ω are shown in Fig. 6.13. It can be easily shown that at $\Omega = -\beta$ and at $\Omega = +\beta$, the function $(\eta^2 - \kappa^2)$ attains its maximum and minimum values respectively.

It should be pointed out that, in general, an atom can execute oscillations corresponding to different resonant frequencies and we have to take into account the various contributions. If $\omega_0, \omega_1, \dots$ represent the resonant frequencies and if f_j represents the fractional number of electrons per unit volume whose resonant frequency is ω_j , Eq. (84) would get modified to the following expression:**

$$n^2 = 1 + \frac{Nq^2}{m\epsilon_0} \sum_j \frac{f_j}{[\omega_j^2 - \omega^2 - 2iK_j\omega]} \quad (92)$$

where K_j represents the damping constant corresponding to the resonant frequency ω_j . Indeed, Eq. (92) describes

*Notice that A is complex; however, if we substitute the expression for A from Eq. (82) in Eq. (81) and take the real part we would get the same expression for $x(t)$ as we had obtained in Sec. 6.4.

**Quantum mechanics also gives a similar result (see, for example, Ref. 6).

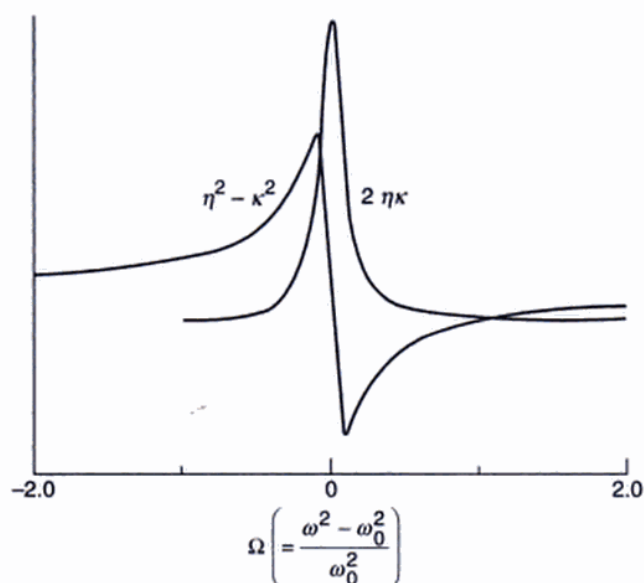


Fig. 6.13 Qualitative variation of $(\eta^2 - \kappa^2)$ and $2\eta\kappa$ with Ω .

correctly the variation of refractive index for most gases. Figure 6.14 shows the dependence of the refractive index of sodium vapour around $\lambda_0 = 5800 \text{ \AA}$. Since D_1 and D_2 lines occur at 5890 \AA and 5896 \AA , one should expect resonant oscillations around these frequencies. This is indeed borne out by the data shown in Fig. 6.14. The variation of the refractive index can be accurately fitted with the formula

$$n^2 = 1 + \frac{A}{v^2 - v_1^2} + \frac{B}{v^2 - v_2^2} \quad (93)$$

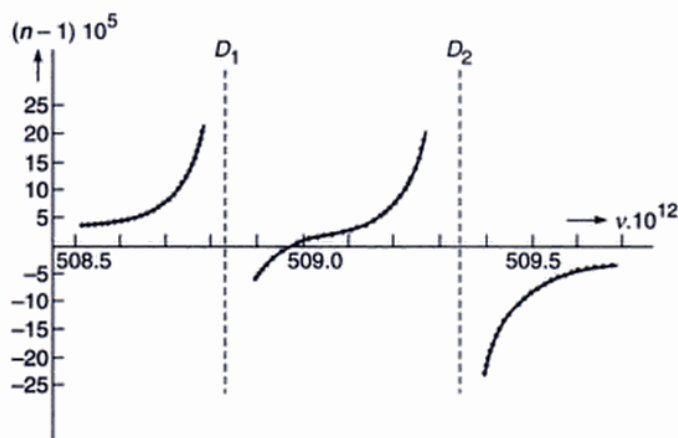


Fig. 6.14 The measured variation of refractive index of sodium with frequency around the D_1 and D_2 lines. The measurements are of Roschdestwensky; the figure has been adapted from Ref. 1.

where we have neglected the presence of damping forces which is justified except when one is very close to the resonance.

It may be worthwhile mentioning that in a liquid, the molecules are very close to one another and the dipoles interact between themselves. If we take this interaction into account, we would get*

$$\frac{n^2 - 1}{n^2 + 2} = \frac{Nq^2}{3m\epsilon_0} \sum_j \frac{f_j}{\omega_j^2 - \omega^2} \quad (94)$$

where we have neglected the presence of damping. For liquids, whose molecules do not have a permanent dipole moment (e.g., H_2 , O_2 , etc.) Eq. (94) gives a fairly accurate description. However, liquids whose molecules possess permanent dipole moments (e.g., H_2O) one has to carry out a different analysis.

6.6 RAYLEIGH SCATTERING

We end this chapter by giving a brief account of Rayleigh scattering. Throughout our analysis we will assume that each scattering center behaves independently—an assumption which will be valid for a gas where the average interatomic spacing is greater than the wavelength.

As discussed in Sec. 6.5, the incident electric field \mathbf{E} produces a dipole moment given by [see Eqs (64) and (65)]

$$\mathbf{p} = \frac{q^2}{m(\omega_0^2 - \omega^2)} \mathbf{E} \quad (95)$$

where ω_0 represents the natural frequency of the atom. To keep the analysis simple, we are neglecting the effect of damping although it can be taken into account without much difficulty. Now, an oscillating dipole given by

$$\mathbf{p} = \mathbf{p}_0 e^{-i\omega t} \quad (96)$$

radiates energy at a rate (see Sec. 20.4.1)

$$\bar{P} = \frac{\omega^4 p_0^2}{12\pi\epsilon_0 c^3} \quad (97)$$

or

$$\bar{P} = \frac{\omega^4}{12\pi\epsilon_0 c^3} \frac{q^4}{m^2 (\omega_0^2 - \omega^2)^2} E_0^2 \quad (98)$$

Thus if N represents the number of atoms per unit volume, then the total energy radiated away (per unit volume) would be $N\bar{P}$.

We assume the electromagnetic wave to be propagating along the x -direction. The intensity of the wave is given by [see Eq. (78) of Chapter 20]

$$I = \frac{1}{2} \epsilon_0 c E_0^2 \quad (99)$$

* See, for example, Ref. 1. Notice that when n is very close to unity (i.e., for a dilute fluid), Eq. (94) reduces to Eq. (92).

Thus the change in the intensity of the electro-magnetic wave as it propagates through a distance dx is given by

$$dI = -N \bar{P} dx$$

or

$$\frac{dI}{I} = -\gamma dx \quad (100)$$

where

$$\gamma = \frac{N \omega^4}{6\pi \epsilon_0 c^4} \frac{q^4}{m^2 (\omega_0^2 - \omega^2)^2} \quad (101)$$

The integration of Eq. (100) is simple:

$$I = I_0 e^{-\gamma x} \quad (102)$$

implying that γ represents the attenuation coefficient. For most atoms ω_0 lies in the ultraviolet region; for example, for the hydrogen atom $\hbar \omega_0 \approx$ few electron volts. Thus if we assume $\omega \ll \omega_0$, then γ becomes proportional to ω^4 or

$$\gamma \propto \frac{1}{\lambda^4} \quad (103)$$

which represents the famous $1/\lambda^4$ Rayleigh scattering law and is responsible for the blue colour of the sky (because it is the blue component which is predominantly scattered). Similarly, the blue component of the light coming from the setting sun is predominantly scattered out resulting in the red colour of the setting sun. Indeed, if the colour of the setting (or rising) sun is deep red, one can infer that the pollution level is high. Now, for a gas,

$$n^2 - 1 = \frac{N q^2}{m \epsilon_0 (\omega_0^2 - \omega^2)} \quad (104)$$

[See Eq. (68)]. For air, since the refractive index is very close to unity, we may write

$$n - 1 \approx \frac{N q^2}{2 m \epsilon_0 (\omega_0^2 - \omega^2)} \quad (105)$$

using which, Eq. (101) can be written in the following convenient form

$$\begin{aligned} \gamma &= \frac{2}{3\pi N} \left(\frac{\omega}{c} \right)^4 (n - 1)^2 \\ &= \frac{2k^4}{3\pi N} (n - 1)^2; \quad k = \frac{\omega}{c} \end{aligned} \quad (106)$$

For air at NTP, the quantity $n - 1 \approx 2.78 \times 10^{-4}$ in the entire region of the visible spectrum. With $N \approx 2.7 \times 10^{19}$ molecules/cm³ we obtain

$$L = \frac{1}{\gamma} = 27 \text{ km, } 128 \text{ km and } 188 \text{ km}$$

for $\lambda = 4000 \text{ \AA}$ (violet), 5900 \AA (yellow) and 6500 \AA (red) respectively. The quantity L represents the distance in which the intensity decreases by a factor of e .

We may conclude this chapter by mentioning that in the 1929 edition of Encyclopaedia Britannica, Lord Rayleigh wrote in an article on SKY:

SKY: *The apparent covering of the atmosphere, the overarching heaven... It is a matter of common observation that the blue of the sky is highly variable, even on days that are free from clouds. The colour usually deepens toward the zenith and also with the elevation of the observer... Closely associated with the colour is the polarization of light from the sky. This takes place in a plane passing through the sun, and attains a maximum about 90° therefrom.*

SUMMARY

- The most fundamental vibration associated with wave motion is the simple harmonic motion.
- When a point rotates on the circumference of a circle with a uniform angular velocity, the foot of the perpendicular on any one of its diameters will execute simple harmonic motion.
- When an external sinusoidal force is applied to a vibrating system, we have what is known as forced vibrations. In steady state, the frequency of the forced vibrations is the same as that of the external force.
- When a lightwave interacts with an atom, we may assume the electrons to behave like oscillators with resonant frequency ω_0 . The electric field of the lightwave polarizes the molecules of the gas, producing oscillating dipole moments from which one can make a first principle calculation of the refractive index to obtain

$$n^2(\omega) \approx 1 + \frac{N q^2}{m \epsilon_0 (\omega_0^2 - \omega^2 - 2iK\omega)}$$

where m is the mass of the electron, q the magnitude of the charge of the electron, N is the number of electrons per unit volume and K is the damping constant. Because of the fact that an oscillating dipole radiates energy, the lightwave gets attenuated; this

leads to the famous $\frac{1}{\lambda^4}$ Rayleigh scattering law which is responsible for the red colour of the rising sun and blue colour of the sky.

PROBLEMS

- 6.1 The displacement in a string is given by the following equation:

$$y(x, t) = a \cos\left(\frac{2\pi}{\lambda} x - 2\pi \nu t\right)$$

where a , λ and ν represent the amplitude, wavelength and the frequency of the wave. Assume $a = 0.1$ cm, $\lambda = 4$ cm, $\nu = 1$ sec⁻¹. Plot the time dependence of the displacement at $x = 0, 0.5$ cm, 1.0 cm, 1.5 cm, 2 cm, 3 cm and 4 cm. Interpret the plots physically.

- 6.2 The displacement associated with a standing wave on a sonometer is given by the following equation:

$$y(x, t) = 2a \sin\left(\frac{2\pi}{\lambda} x\right) \cos 2\pi \nu t$$

If the length of the string is L then the allowed values of λ are $2L, 2L/2, 2L/3, \dots$ (see Sec. 11.2). Consider the case when $\lambda = 2L/5$; study the time variation of displacement in each loop and show that alternate loops vibrate in phase (with different points in a loop having different amplitudes) and adjacent loops vibrate out of phase.

- 6.3 A tunnel is dug through the earth as shown in Fig. 6.15. A mass is dropped at the point A along the tunnel. Show that it will execute simple harmonic motion. What will the time period be?

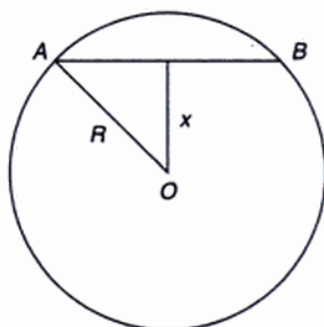


Fig. 6.15 For Problem 6.3.

- 6.4 A 1 g mass is suspended from a vertical spring. It executes simple harmonic motion with period 0.1 sec. By how much distance had the spring stretched when the mass was attached?
- 6.5 A stretched string is given simultaneous displacement in the x - and y - directions such that

$$x(z, t) = a \cos\left(\frac{2\pi}{\lambda} z - 2\pi \nu t\right)$$

and

$$y(z, t) = a \cos\left(\frac{2\pi}{\lambda} z - 2\pi \nu t\right)$$

Study the resultant displacement (at a particular value of z) as a function of time.

- 6.6 In Problem 6.5, if

$$x(z, t) = a \cos\left(\frac{2\pi}{\lambda} z - 2\pi \nu t\right)$$

and

$$y(z, t) = a \sin\left(\frac{2\pi}{\lambda} z - 2\pi \nu t\right)$$

what will be the resultant displacement?

- 6.7 As mentioned in Sec. 6.5, alkali metals are transparent to ultraviolet light. Assuming that the refractive index is primarily due to the free electrons and that there is one free electron per atom, calculate

$$\lambda_p \left(= \frac{2\pi c}{\omega_p} \right) \text{ for Li, K and Rb. You may assume that}$$

the atomic weights of Li, K and Rb are 6.94, 39.10 and 85.48 respectively and that the corresponding densities are 0.534, 0.870 and 1.532 g/cm³. Also, the values of various physical constants are: $m = 9.109 \times 10^{-31}$ kg, $q = 1.602 \times 10^{-19}$ C and $\epsilon_0 = 8.854 \times 10^{-12}$ C/N-m².

[Ans: 1550 Å, 2884 Å and 3214 Å; the corresponding experimental values are 1551 Å, 3150 Å and 3400 Å respectively].

- 6.8 (a) In a metal, the electrons can be assumed to be essentially free. The drift velocity of the electron satisfies the following equation

$$m \frac{dv}{dt} + m \nu v = F = -q E_0 e^{-i\omega t}$$

where ν represents the collision frequency. Calculate the steady state current density ($\mathbf{J} = -Nq\mathbf{v}$) and show that the conductivity is given by

$$\sigma(\omega) = \frac{Nq^2}{m} \frac{1}{\nu - i\omega}$$

- (b) If \mathbf{r} represents the displacement of the electron, show that

$$\mathbf{P} = -Nq\mathbf{r} = - \frac{Nq^2}{m(\omega^2 + i\omega\nu)} \mathbf{E}$$

which represents the polarization. Using the above equation show that

$$\kappa(\omega) = 1 - \frac{Nq^2}{m\epsilon_0(\omega^2 + i\omega\nu)}$$

which represents the dielectric constant variation for a free-electron gas.

- 6.9** Assuming that each atom of copper contributes one free electron and that the low frequency conductivity σ is about 6×10^7 mhos/metre, show that $\nu \approx 4 \times 10^{13} \text{ s}^{-1}$. Using this value of ν , show that the conductivity is almost real for $\omega \leq 10^{11} \text{ s}^{-1}$. For $\omega = 10^8 \text{ s}^{-1}$ calculate the complex dielectric constant and compare its value with the one obtained for infrared frequencies.

It may be noted that for small frequencies, only one of the electrons of a copper atom can be considered to be free. On the other hand, for X-ray frequencies all the electrons may be assumed to be free (see Problems 6.10, 6.11 and 6.12). Discuss the validity of the above argument.

- 6.10** Show that for high frequencies ($\omega \gg \nu$) the dielectric constant (as derived in Problem 6.8) is essentially real with frequency dependence of the form

$$\kappa = 1 - \frac{\omega_p^2}{\omega^2}$$

where $\omega_p = \left(\frac{N q^2}{m \epsilon_0} \right)^{1/2}$ is known as the plasma frequency. The above dielectric constant variation is indeed valid for X-ray wavelengths in many metals. Assuming that at such frequencies all the electrons can be assumed to be free, calculate ω_p for copper for which the atomic number is 29, mass number is 63, and density is 9 g/cm^3 .

[Ans: $\sim 9 \times 10^{16} \text{ sec}^{-1}$]

- 6.11** Obtain an approximate value for the refractive index of metallic sodium corresponding to $\lambda = 1 \text{ \AA}$. Assume all the electrons of sodium to be free.
- 6.12** In an ionic crystal (like NaCl, CaF₂, etc.), one has to take into account infra-red resonance oscillations of the ions and Eq. (68) modifies to

$$n^2 = 1 + \frac{N q^2}{m \epsilon_0 (\omega_1^2 - \omega^2)} + \frac{p N q^2}{M \epsilon_0 (\omega_2^2 - \omega^2)}$$

where M represents the reduced mass of the two ions and p represents the valency of the ion ($p = 1$ for Na⁺, Cl⁻; $p = 2$ for Ca⁺⁺, F₂⁻). Show that the above equation can be written in the form

$$n^2 = n_\infty^2 + \frac{A_1}{\lambda^2 - \lambda_1^2} + \frac{A_2}{\lambda^2 - \lambda_2^2}$$

where

$$\lambda_1 = \frac{2\pi c}{\omega_1}, \quad \lambda_2 = \frac{2\pi c}{\omega_2}$$

$$A_1 = \frac{N q^2}{4\pi^2 c^2 \epsilon_0 m} \lambda_1^4, \quad A_2 = \frac{p N q^2}{4\pi^2 c^2 \epsilon_0 M} \lambda_2^4$$

- 6.13** The refractive index variation for CaF₂ (in the visible region of the spectrum) can be written in the form*

$$n^2 = 6.09 + \frac{6.12 \times 10^{-15}}{\lambda^2 - 8.88 \times 10^{-15}} + \frac{5.10 \times 10^{-9}}{\lambda^2 - 1.26 \times 10^{-9}}$$

where λ is in metres

- (a) Plot the variation of n^2 with λ in the visible region.
- (b) From the values of A_1 and A_2 show that $m/M \approx 2.07 \times 10^{-5}$ and compare this with the exact value.
- (c) Show that the value of n_∞ obtained by using the constants A_1 , A_2 , λ_1 and λ_2 agrees reasonably well with the experimental value.
- 6.14** (a) The refractive index of a plasma (neglecting collisions) is approximately given by (see Sec. 6.6)

$$n^2 = 1 - \frac{\omega_p^2}{\omega^2}$$

where

$$\omega_p = \left(\frac{N q^2}{m \epsilon_0} \right)^{1/2} \approx 56.414 N^{1/2} \text{ s}^{-1}$$

is known as the plasma frequency. In the ionosphere, the maximum value of N_0 is $\approx 10^{10} - 10^{12}$ electrons/m³. Calculate the plasma frequency. Notice that at high frequencies $n^2 \approx 1$; thus high frequency waves (like the one used in TV) are not reflected by the ionosphere. On the other hand, for low frequencies, the refractive index is imaginary (like in a conductor—see Sec. 21.3) and the beam gets reflected. This fact is used in long distance radio communications (See Fig. 2.20).

- (b) Assume that for $x \approx 200 \text{ km}$, $N = 10^{12}$ electrons/m³ and that the electron density increases to 2×10^{12} electrons/m³ at $x \approx 300 \text{ km}$. For $x \geq 300 \text{ km}$, the electron density decreases. Assuming a parabolic variation of N , plot the corresponding refractive index variation.

*Quoted from Ref. 9; measurements are of Paschen.

REFERENCES AND SUGGESTED READINGS

1. C.J.F. Bottcher, *Theory of Electric Polarization*, Elsevier Publishing Co., Amsterdam, 1952.
2. H.J.J. Braddick, *Vibrations, Waves and Diffraction*, McGraw-Hill Publishing Co., London, 1965.
3. F.S. Crawford, *Waves and Oscillations: Berkeley Physics Course*, Vol. III, McGraw-Hill Book Co., New York, 1968.
4. R.P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Mass., 1965.
5. A.P. French, *Vibrations and Waves*, Arnold-Heineman India, New Delhi, 1973.
6. R. Loudon, *The Quantum Theory of Light*, Clarendon Press, Oxford, 1973.
7. H.J. Pain, *The Physics of Vibrations and Waves*, John Wiley & Sons, London, 1968.
8. R. Resnick and D. Halliday, *Physics*, Part I, John Wiley & Sons, New York, 1966.
9. A. Sommerfeld, *Optics*, Academic Press, New York, 1964.
10. J.M. Stone, *Radiation and Optics*, McGraw-Hill Book Co., New York, 1963.

Chapter 7

Fourier Series and Applications

... Reimann (in one of his publications in 1867) asserts that when Fourier, in his first paper to the Paris Academy in 1807, stated that a completely arbitrary function could be expressed in such a series, his statement so surprised Lagrange that he denied possibility in the most definite terms. It should also be noted that he (Fourier) was the first to allow that the arbitrary function might be given by different analytical expressions in different parts of the interval ...

H.A. Carslaw (1930)

7.1 FOURIER ANALYSIS

Fourier series and Fourier integrals are extensively used in the theory of vibrations and waves. As such, we devote this chapter to the study of Fourier series and Fourier integrals. The results obtained will be used in subsequent chapters. Now, according to Fourier's theorem, any periodic vibration can be expressed as a sum of the sine and cosine functions whose frequencies increase in the ratio of natural numbers. Thus, a periodic function with period T , i.e.,

$$f(t + nT) = f(t); \quad n = 0, \pm 1, \pm 2, \dots; \quad (1)$$

can be expanded in the form:

$$\begin{aligned} f(t) &= \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2n\pi}{T} t\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2n\pi}{T} t\right) \\ &= \frac{1}{2} a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega t) + \sum_{n=1}^{\infty} b_n \sin(n\omega t) \end{aligned} \quad (2)$$

where

$$\omega = \frac{2\pi}{T} \quad (3)$$

represents the fundamental frequency. Actually, for the expansion to be possible, the function $f(t)$ must satisfy certain conditions. The conditions are that the function $f(t)$ in one period (i.e., in the interval $t_0 < t < t_0 + T$) must be (a) single valued, (b) piecewise continuous (i.e., it can have at most a finite number of finite discontinuities) and (c) can have only a finite number of maxima and minima. These conditions are known as Dirichlet's conditions and are almost always satisfied in all problems that one encounters in physics.

The coefficients a_n and b_n can easily be determined by using the following properties of the trigonometric functions:

$$\int_{t_0}^{t_0+T} \cos n\omega t \cos m\omega t dt = \begin{cases} 0 & \text{if } m \neq n \\ T/2 & \text{if } m = n \end{cases} \quad (4)$$

$$\int_{t_0}^{t_0+T} \sin n\omega t \sin m\omega t dt = \begin{cases} 0 & \text{if } m \neq n \\ T/2 & \text{if } m = n \end{cases} \quad (5)$$

$$\int_{t_0}^{t_0+T} \sin n\omega t \cos m\omega t dt = 0 \quad (6)$$

The above equations can easily be derived. For example, for $m = n$,

$$\begin{aligned} \int_{t_0}^{t_0+T} \cos n\omega t \cos m\omega t dt &= \int_{t_0}^{t_0+T} \cos^2 n\omega t dt \\ &= \frac{1}{2} \int_{t_0}^{t_0+T} [1 + \cos 2n\omega t] dt = \frac{T}{2} \end{aligned}$$

Similarly, for $m \neq n$

$$\begin{aligned} \int_{t_0}^{t_0+T} \cos n\omega t \cos m\omega t dt \\ = \frac{1}{2} \int_{t_0}^{t_0+T} [\cos(n-m)\omega t + \cos(n+m)\omega t] dt \end{aligned}$$

$$= \frac{1}{2} \left[\frac{1}{(n-m)\omega} \sin(n-m)\omega t + \frac{1}{(n+m)\omega} \sin(n+m)\omega t \right]_{t_0}^{t_0+T} = 0$$

In order to determine the coefficients a_n and b_n we first multiply Eq. (2) by dt and integrate from t_0 to $t_0 + T$:

$$\int_{t_0}^{t_0+T} f(t) dt = \frac{1}{2} a_0 \int_{t_0}^{t_0+T} dt + \sum_{n=1}^{\infty} a_n \int_{t_0}^{t_0+T} \cos n\omega t dt + \sum_{n=1}^{\infty} b_n \int_{t_0}^{t_0+T} \sin n\omega t dt = \frac{T}{2} a_0$$

where we have used Eqs (4) and (6) for $m = 0$. Thus

$$a_0 = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) dt \quad (7)$$

Next, if we multiply Eq. (2) by $\cos(m\omega t) dt$ and integrate from t_0 to $t_0 + T$ we would obtain

$$\begin{aligned} \int_{t_0}^{t_0+T} f(t) \cos(m\omega t) dt &= \frac{1}{2} a_0 \int_{t_0}^{t_0+T} \cos(m\omega t) dt + \\ &\sum_{n=1}^{\infty} a_n \int_{t_0}^{t_0+T} \cos(m\omega t) \cos(n\omega t) dt + \\ &\sum_{n=1}^{\infty} b_n \int_{t_0}^{t_0+T} \cos(m\omega t) \sin(n\omega t) dt \\ &= \frac{T}{2} a_m \end{aligned}$$

where we have used Eqs. (4) and (6). We may combine the above equation with Eq. (7) to write

$$a_n = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \cos n\omega t dt; \quad n = 0, 1, 2, 3, \dots \quad (8)$$

Similarly,

$$b_n = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \sin n\omega t dt; \quad n = 1, 2, 3, \dots \quad (9)$$

It should be pointed out that the value of t_0 is quite arbitrary. In some problems it is convenient to choose

$$t_0 = -T/2$$

then

$$a_n = \frac{2}{T} \int_{-T/2}^{+T/2} f(t) \cos n\omega t dt; \quad n = 0, 1, 2, \dots$$

and

$$b_n = \frac{2}{T} \int_{-T/2}^{+T/2} f(t) \sin n\omega t dt; \quad n = 0, 1, 2, \dots$$

Such a choice is particularly convenient when the function is even (i.e., $f(t) = f(-t)$) or odd (i.e., $f(t) = -f(-t)$). In the former case $b_n = 0$ whereas in the latter case $a_n = 0$. In some problems, it is convenient to choose $t_0 = 0$.

Example 7.1 Consider a periodic function of the form

$$\left. \begin{aligned} f(t) &= t \quad \text{for } -\tau < t < +\tau \\ f(t + 2n\tau) &= f(t) \end{aligned} \right\} \quad (10)$$

(see Fig. 7.1). Such a function is referred to as a saw tooth function. In this example, we will expand the above function in a Fourier series. Now, since $f(t)$ is an odd function of t , $a_n = 0$ and

$$\begin{aligned} b_n &= \frac{2}{T} \int_{-\tau}^{+\tau} f(t) \sin(n\omega t) dt \\ &= \frac{1}{\tau} \int_{-\tau}^{+\tau} t \sin(n\omega t) dt \end{aligned}$$

Notice that the periodicity is 2τ and, therefore, $\omega = \pi/\tau$. Carrying out the integration we obtain

$$b_n = \frac{2}{\tau} \left[-\frac{t}{n\omega} \cos n\omega t + \frac{1}{n\omega} \left\{ \frac{1}{n\omega} \sin n\omega t \right\} \right]_0^{\tau}$$

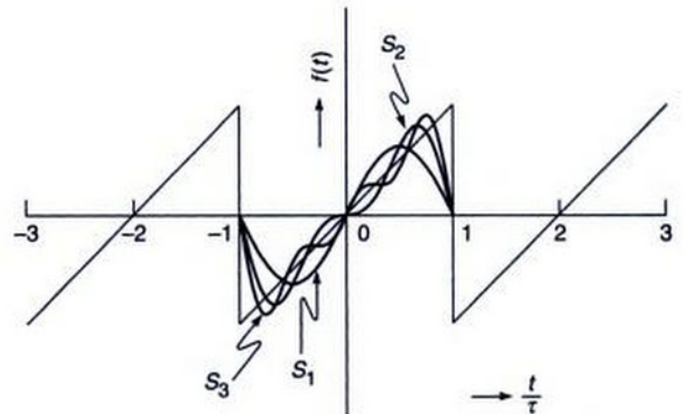


Fig. 7.1 The saw tooth function; S_1 , S_2 and S_3 represent the partial sums corresponding to the saw tooth function.

$$= -\frac{2\tau}{n\pi} \cos n\pi = (-1)^{n+1} \frac{2\tau}{n\pi} \quad (11)$$

Thus

$$\begin{aligned} f(t) &= \frac{2\tau}{\pi} \sum_{n=1,2,\dots} \frac{(-1)^{n+1}}{n} \sin n\omega t \\ &= \frac{2\tau}{\pi} \left[\sin \omega t - \frac{1}{2} \sin 2\omega t + \frac{1}{3} \sin 3\omega t - \dots \right] \quad (12) \end{aligned}$$

In Fig. 7.1 we have also plotted the partial sums which are given by

$$\begin{aligned} S_1 &= \frac{2\tau}{\pi} \sin \omega t; \quad S_2 = \frac{2\tau}{\pi} \left[\sin \omega t - \frac{1}{2} \sin 2\omega t \right] \\ S_3 &= \frac{2\tau}{\pi} \left[\sin \omega t - \frac{1}{2} \sin 2\omega t + \frac{1}{3} \sin 3\omega t \right] \end{aligned}$$

It can be seen from the figure that as n increases, the sum S_n approaches the function $f(t)$.

Example 7.2 In this example, we will Fourier expand the function defined by the following equations:

$$\begin{aligned} f(t) &= -A \quad \text{for} \quad -\frac{T}{2} < t < 0 \\ &= +A \quad \text{for} \quad 0 < t < t + \frac{T}{2} \quad (13) \end{aligned}$$

and

$$f(t+T) = f(t)$$

The function is plotted in Fig. 7.2. Once again the function is an odd function; consequently $a_n = 0$ and

$$\begin{aligned} b_n &= \frac{2}{T} \int_0^{T/2} A \sin(n\omega t) dt = \frac{4A}{T} \frac{1}{n\omega} [-\cos n\omega t]_0^{T/2} \\ &= \frac{2A}{n\pi} [1 - \cos n\pi] = \frac{2A}{n\pi} [1 - (-1)^n] \end{aligned}$$

Thus

$$\begin{aligned} f(t) &= \frac{2A}{\pi} \sum_{n=1,2,3,\dots} \frac{1}{n} [1 - (-1)^n] \sin n\omega t \\ &= \frac{4A}{\pi} \left[\sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t + \dots \right] \end{aligned}$$

The partial sums

$$\begin{aligned} S_1 &= \frac{4A}{\pi} \sin \omega t; \quad S_2 = \frac{4A}{\pi} \left(\sin \omega t + \frac{1}{3} \sin 3\omega t \right) \\ S_3 &= \frac{4A}{\pi} \left[\sin \omega t + \frac{1}{3} \sin 3\omega t + \frac{1}{5} \sin 5\omega t \right] \end{aligned}$$

are also plotted in Fig. 7.2.

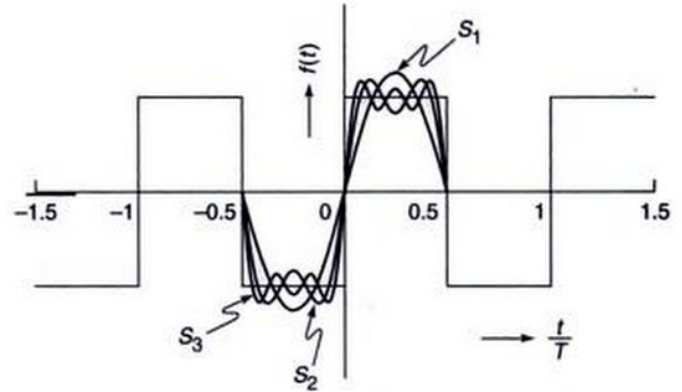


Fig. 7.2 A plot of the periodic step function defined by Eq. (13). S_1 , S_2 and S_3 represent the corresponding partial sums.

7.2 TRANSVERSE VIBRATIONS OF A PLUCKED STRING

An interesting application of the Fourier series lies in studying the transverse vibrations of a plucked string.

Let us consider a stretched string, fixed at the two ends A and B . One of the ends (A) is chosen as the origin. In the equilibrium position of the string, it is assumed to lie along the x -axis (see Fig. 7.3). A point of the string is moved upwards by a distance d ; the corresponding shape of the string is shown as dashed line in Fig. 7.3. If the displacement occurs at a distance a from the origin, the equation of the string (in its displaced position) would be given by the following equation:

$$\begin{aligned} y &= \frac{d}{a} x \quad \text{for} \quad 0 < x < a \\ &= \frac{d}{L-a} (L-x) \quad \text{for} \quad a < x < L \quad (15) \end{aligned}$$

where L represents the length of the string. Now, if the string is released from this position at $t = 0$, we would like to determine the shape of the string at any subsequent time.



Fig. 7.3 The plucked string; AB represents the equilibrium position. The dashed lines show the displaced position at $t = 0$.

We will show in Sec. 9.6 that the displacement $y(x, t)$ satisfies the following wave equation:

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 y}{\partial t^2} \quad (16)$$

where $v (= \sqrt{T/\rho})$ represents the speed of the transverse waves, T being the tension in the string and ρ , the mass per unit length. We would like to solve Eq. (16) subject to the following boundary conditions:

$$(a) \ y = 0 \text{ at } x = 0 \text{ and } x = L \text{ for all values of } t \quad (17)$$

$$(b) \text{ At } t = 0$$

$$(i) \ \frac{\partial y}{\partial t} = 0 \quad \text{for all values of } x. \quad (18)$$

$$(ii) \ y(x, t = 0) = \frac{d}{a} x \quad \text{for } 0 < x < a$$

$$= \frac{d}{L-a} (L-x) \quad \text{for } a < x < L \quad (19)$$

Assuming a time dependence of the form $\cos \omega t$ (or $\sin \omega t$):

$$y(X, t) = X(x) \cos \omega t$$

we obtain

$$\frac{d^2 X}{dx^2} = -\frac{\omega^2}{v^2} X(x)$$

or,

$$\frac{d^2 X}{dx^2} + k^2 X(x) = 0 \quad (20)$$

where

$$k = \frac{\omega}{v} \quad (21)$$

*Rigorously we should proceed by using the method of separation of variables; thus we assume

$$y(x, t) = X(x) T(t)$$

where $X(x)$ is a function of x alone and $T(t)$ is a function of t alone. Substituting in Eq. (16), we get

$$\frac{1}{X(x)} \frac{d^2 X}{dx^2} = \frac{1}{v^2} \frac{1}{T(t)} \frac{d^2 T}{dt^2} = -k^2$$

Since the term $\frac{1}{X} \frac{d^2 X}{dx^2}$ is a function of x alone and the term $\frac{1}{v^2} \frac{1}{T} \frac{d^2 T}{dt^2}$ is a function of t alone, each term must be equal to a constant which we have put equal to $-k^2$. Thus

$$\frac{d^2 T}{dt^2} + \omega^2 T(t) = 0$$

and

$$\frac{d^2 X}{dx^2} + k^2 X(x) = 0$$

where

$$\omega = kv$$

The solution of Eq. (20) is simple*:

$$X(x) = A \sin kx + B \cos kx \quad (22)$$

Thus

$$y(x, t) = (A \sin kx + B \cos kx) (C \cos \omega t + D \sin \omega t)$$

Now

$$y(x, t)|_{x=0} = 0 \quad \text{for all values of } t$$

Thus $B = 0$ and we obtain

$$y(x, t) = \sin kx [C \cos \omega t + D \sin \omega t]$$

where we have absorbed A in C and D . Since

$$y(x, t)|_{x=L} = 0 \quad (\text{for all values of } t)$$

we must have

$$\sin kL = 0$$

or

$$kL = n\pi; \ n = 1, 2, 3, \dots \quad (23)$$

Thus, only discrete values of k (and hence of ω) are permissible; these are given by

$$k_n = \frac{n\pi}{L}; \ n = 1, 2, \dots \quad (24)$$

giving

$$\omega_n = \frac{n\pi v}{L}; \ n = 1, 2, \dots \quad (25)$$

Equation (25) gives the frequencies of the various modes of the string. The mode corresponding to the lowest frequency ($n = 1$) is known as the fundamental mode.

Thus the solution of Eq. (16) satisfying the boundary condition given by Eq. (17) would be given by

$$y(x, t) = \sum_{n=1,2,3,\dots} \sin k_n x [C_n \cos \omega_n t + D_n \sin \omega_n t] \quad (26)$$

Differentiating partially with respect to t , we get

$$\begin{aligned} \left. \frac{\partial y}{\partial t} \right|_{t=0} &= \sum_n \sin k_n x [-\omega_n C_n \sin \omega_n t + \omega_n D_n \cos \omega_n t] \Big|_{t=0} \\ &= \sum_n \omega_n D_n \sin k_n x \end{aligned} \quad (27)$$

Since

$$\left. \frac{\partial y}{\partial t} \right|_{t=0} = 0 \quad \text{for all values of } x$$

we must have

$$D_n = 0 \quad \text{for all } n$$

Thus

$$y(x, t) = \sum_{n=1,2,3,\dots} C_n \sin k_n x \cos \omega_n t \quad (28)$$

or

$$y(x, 0) = \sum_n C_n \sin \left(\frac{n\pi}{L} x \right) \quad (29)$$

The above equation is essentially a Fourier series and in order to determine C_n , we multiply both sides of Eq. (29) by

$\sin \left(\frac{m\pi}{L} x \right) dx$ and integrate from 0 to L to obtain:

$$C_m = \frac{2}{L} \int_0^L y(x, 0) \sin \left(\frac{m\pi}{L} x \right) dx \quad (30)$$

where we have used the relation

$$\int_0^L \sin \frac{n\pi x}{L} \sin \frac{m\pi x}{L} dx = \begin{cases} 0 & \text{if } m \neq n \\ L/2 & \text{if } m = n \end{cases} \quad (31)$$

[cf. Eq. (5)]. Substituting the expression for $y(x, 0)$ from Eq. (19), we obtain

$$\begin{aligned} C_n &= \frac{2}{L} \left[\frac{d}{a} \int_0^a x \sin \left(\frac{n\pi}{L} x \right) dx + \right. \\ &\quad \left. \frac{d}{L-a} \int_a^L (L-x) \sin \left(\frac{n\pi}{L} x \right) dx \right] \\ &= \frac{2dL^2}{a(L-a)\pi^2 n^2} \sin \left(\frac{n\pi}{L} a \right) \end{aligned}$$

On substituting in Eq. (28), we finally obtain

$$y(x, t) = \frac{2dL^2}{a(L-a)\pi^2} \sum_{n=1,2,3,\dots} \frac{1}{n^2} \sin \left(\frac{n\pi}{L} a \right) \times \sin \left(\frac{n\pi}{L} x \right) \cos \left(\frac{n\pi v}{L} t \right) \quad (32)$$

Equation (32) can be used to determine the shape of the string at an arbitrary time t . If the string is plucked at the centre (i.e., $a = L/2$), terms corresponding to $n = 2, 4, 6, \dots$ are absent (i.e., the even harmonics are absent) and Eq. (32) simplifies to

$$y(x, t) = \frac{8d}{\pi^2} \sum_m (-1)^{m+1} \frac{1}{(2m-1)^2} \sin \frac{(2m-1)\pi x}{L} \times \cos \frac{(2m-1)(\pi v t)}{L} \quad (33)$$

7.3 APPLICATION OF FOURIER SERIES IN FORCED VIBRATIONS

Let us consider the forced vibrations of a damped oscillator. The equation of motion would be

$$m \frac{d^2 y}{dt^2} + \Gamma \frac{dy}{dt} + k_0 y = F(t) \quad (34)$$

where Γ represents the damping constant (see Sec. 6.3) and F represents the external force. It has been shown in Sec. 6.4 that if $\Gamma > 0$ and

$$F(t) = F_0 \cos (pt + \theta) \quad (35)$$

then the steady state solution of Eq. (34) is a simple harmonic motion with the frequency of the external force. If $F(t)$ is not a sine or cosine function, a general solution of Eq. (34) is difficult to obtain; however, if $F(t)$ is periodic then we can apply Fourier's theorem to obtain a solution of Eq. (34). For example, let

$$F(t) = \alpha t \quad \text{for } -\tau < t < \tau \quad (36)$$

and

$$F(t + 2n\tau) = F(t); \quad n = 1, 2, \dots$$

The Fourier expansion of such a function was discussed in Example 7.1 and is of the form:

$$\begin{aligned} F(t) &= \sum_n F_n \sin n\omega t \\ &= \frac{2\alpha\tau}{\pi} \left[\sum_{n=1,2,\dots} \frac{(-1)^{n+1}}{n} \sin n\omega t \right] \end{aligned} \quad (37)$$

We next consider the solution of the differential equation

$$m \frac{d^2 y_n}{dt^2} + \Gamma \frac{dy_n}{dt} + k_0 y_n = F_n \sin n\omega t$$

or

$$\frac{d^2 y_n}{dt^2} + K \frac{dy_n}{dt} + \omega_0^2 y_n = A_n \sin n\omega t \quad (38)$$

where

$$K \equiv \frac{\Gamma}{m}; \quad \omega_0^2 \equiv \frac{k_0}{m}$$

and

$$A_n = \frac{F_n}{m} = \frac{(-1)^{n+1}}{n} \frac{2\alpha\tau}{\pi m} \quad (39)$$

The steady state solution of Eq. (38) will be of the form

$$y_n = C_n \sin n\omega t + D_n \cos n\omega t$$

and the solution of Eq. (34) will be of the form

$$y = \sum_n y_n \quad (40)$$

In order to determine C_n and D_n we substitute the above solution in Eq. (38) to obtain

$$\begin{aligned} & -n^2 \omega^2 [C_n \sin n\omega t + D_n \cos n\omega t] \\ & + n\omega K [C_n \cos n\omega t - D_n \sin n\omega t] \\ & + \omega_0^2 [C_n \sin n\omega t + D_n \cos n\omega t] = A_n \sin n\omega t \end{aligned}$$

Thus

$$\text{and} \quad \left. \begin{aligned} (\omega_0^2 - n^2 \omega^2) C_n - n\omega K D_n &= A_n \\ (\omega_0^2 - n^2 \omega^2) D_n - n\omega K C_n &= 0 \end{aligned} \right\} \quad (41)$$

Solving the above equations, we get

$$D_n = - \frac{n\omega K}{(\omega_0^2 - n^2 \omega^2)^2 + n^2 \omega^2 K^2} A_n$$

and

$$C_n = \frac{\omega_0^2 - n^2 \omega^2}{(\omega_0^2 - n^2 \omega^2)^2 + n^2 \omega^2 K^2} A_n$$

Thus the steady state solution can be written in the form

$$y = \sum_n G_n \sin(n\omega t + \theta_n) \quad (42)$$

where the amplitude G_n is given by

$$\begin{aligned} G_n &= (C_n^2 + D_n^2)^{1/2} \\ &= \frac{A_n}{[(\omega_0^2 - n^2 \omega^2)^2 + n^2 \omega^2 K^2]^{1/2}} \end{aligned} \quad (43)$$

7.4 THE FOURIER INTEGRAL

In Sec. 7.1 we had shown that a periodic function can be expanded in the form

$$f(t) = \frac{1}{2} a_0 + \sum_{n=1}^{\infty} [a_n \cos n\omega t + b_n \sin n\omega t] \quad (44)$$

where

$$a_n = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \cos n\omega t \, dt \quad (45)$$

$$b_n = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \sin n\omega t \, dt \quad (46)$$

and

$$T = \frac{2\pi}{\omega} \quad (47)$$

On substituting the above expressions for a_n and b_n in Eq. (44) we get [we must replace t by t' in Eqs. (45) and (46)]:

$$\begin{aligned} f(t) &= \frac{1}{T} \int_{-T/2}^{+T/2} f(t') \, dt' + \\ &\sum_{n=1}^{\infty} \left[\frac{2}{T} \cos n\omega t \int_{-T/2}^{+T/2} f(t') \cos n\omega t' \, dt' \right. \\ &\quad \left. + \frac{2}{T} \sin n\omega t \int_{-T/2}^{+T/2} f(t') \sin n\omega t' \, dt' \right] \end{aligned} \quad (48)$$

or

$$\begin{aligned} f(t) &= \frac{1}{2\pi} \Delta s \int_{-\pi/\Delta s}^{+\pi/\Delta s} f(t') \, dt' + \\ &\sum_{n=1}^{\infty} \frac{\Delta s}{\pi} \int_{-\pi/\Delta s}^{+\pi/\Delta s} f(t') \cos [n\Delta s(t-t')] \, dt' \end{aligned} \quad (49)$$

where

$$\Delta s \equiv \frac{2\pi}{T} = \omega$$

We let $T \rightarrow \infty$ so that $\Delta s \rightarrow 0$; notice that when $T \rightarrow \infty$, the function is no more periodic. Thus if the integral

$$\int_{-\infty}^{+\infty} |f(t')| \, dt'$$

exists (i.e., if it has a finite value) then the first term on the RHS of Eq. (49) would go to zero. Further, since

$$\int_0^{\infty} F(s) ds = \lim_{\Delta s \rightarrow 0} \sum_{n=1}^{\infty} F(n \Delta s) \Delta s \quad (50)$$

we have

$$f(t) = \frac{1}{\pi} \int_0^{\infty} \left[\int_{-\infty}^{+\infty} f(t') \cos[s(t-t')] dt' \right] ds \quad (51)$$

Equation (51) is known as the Fourier integral. Since the cosine function inside the integral is an even function of s , we may write

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} f(t') \cos[s(t-t')] dt' \right] ds \quad (52)$$

Further, since $\sin[s(t-t')]$ is an odd function of s ,

$$\frac{i}{2\pi} \int_{-\infty}^{+\infty} \left[\int_{-\infty}^{+\infty} f(t') \sin[s(t-t')] dt' \right] ds = 0 \quad (53)$$

If we add (or subtract) the above two equations, we will get

$$f(t) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} f(t') e^{\pm i\omega(t-t')} dt' d\omega \quad (54)$$

where we have replaced s by ω . Equation (54) is usually referred to as the Fourier integral theorem. Thus, if

$$F(\omega) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{\pm i\omega t} dt \quad (55)$$

then

$$f(t) \equiv \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\omega) e^{\mp i\omega t} d\omega \quad (56)$$

The function $F(\omega)$ is known as the Fourier transform of $f(t)$. For a time dependent function $f(t)$, $F(\omega)$ is usually referred to as its frequency spectrum. Equations (55) and (56) are also written in the form

$$F(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \quad (57)$$

with

$$f(t) = \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \quad (58)$$

In the next chapter we will use the above representation to study the propagation of optical pulses in a dispersive medium (see Sec. 8.3). We can also write

$$G(k) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(x) e^{-ikx} dx \quad (59)$$

with

$$f(x) = \int_{-\infty}^{+\infty} G(k) e^{+ikx} dk \quad (60)$$

where k is often referred to as spatial frequency—a concept that is extensively used in Fourier Optics^{6,7} [see Sec. 16.11].

Example 7.3 As an example, we consider the Fourier transform of the Gaussian function (see Fig. 7.4).

$$f(t) = A e^{-t^2/t_0^2} \quad (61)$$

Thus the Fourier transform is given by [using Eq. (57)]

$$\begin{aligned} F(\omega) &= \frac{A}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{t_0^2}} e^{-i\omega t} dt \\ &= \frac{A t_0}{2\sqrt{\pi}} e^{-\frac{\omega^2 t_0^2}{4}} \end{aligned} \quad (62)$$

where we have used the integral (see Appendix A):

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2 + \beta x} dx = \sqrt{\frac{\pi}{\alpha}} e^{\beta^2/4\alpha} \quad (63)$$

The function $F(\omega)$ [as given by Eq. (62)] is also plotted in Fig. 7.4. Thus the Fourier transform of a Gaussian is a Gaussian. We denote the full width at half maximum (usually abbreviated as FWHM) of $f(t)$ by Δt ; thus at $t = \pm \frac{1}{2} \Delta t$, the function $f(t)$ attains half of its maximum value:

$$\frac{1}{2} A = A \exp \left[-\frac{(\Delta t)^2}{4 t_0^2} \right]$$

Thus

$$\Delta t = 2 \sqrt{\ln 2} t_0 \approx 1.67 t_0$$

Similarly, if $\Delta \omega$ denotes the FWHM of $F(\omega)$ then (see Fig. 7.4)

$$\Delta \omega = \frac{4 \sqrt{\ln 2}}{t_0} = \frac{3.34}{t_0}$$

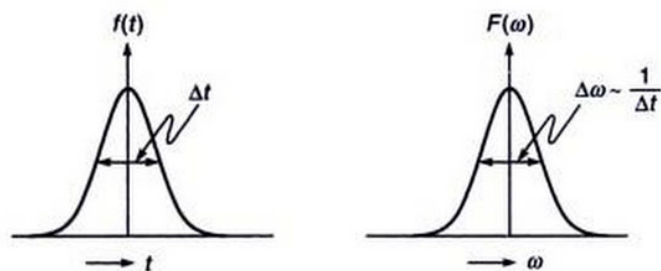


Fig. 7.4 The Fourier transform of a Gaussian temporal function is a Gaussian function in the frequency space.

From the above two equations we get the uncertainty relation

$$\Delta\omega \Delta t \sim 1 \quad (64)$$

which is in general, always valid. We will discuss more examples in Chapters 8 and 15.

SUMMARY

- A periodic function with period T , i.e.
 $f(t + nT) = f(t); \quad n = 0, \pm 1, \pm 2, \dots$
 can be expanded in the form

$$\begin{aligned} f(t) &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{2\pi n}{T}t\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{2\pi n}{T}t\right) \\ &= \frac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos(n\omega t) + \sum_{n=1}^{\infty} b_n \sin(n\omega t) \end{aligned}$$

where

$$\omega = \frac{2\pi}{T}$$

represents the fundamental frequency. The above infinite series is known as the Fourier series and the coefficients a_n and b_n are given by

$$a_n = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \cos n\omega t; \quad n = 0, 1, 2, 3, \dots$$

and

$$b_n = \frac{2}{T} \int_{t_0}^{t_0+T} f(t) \sin n\omega t; \quad n = 0, 1, 2, 3, \dots$$

- Transverse vibrations of a plucked string and forced vibrations can be studied by using Fourier series.
- For a time dependent function $f(t)$, its Fourier transform is defined by the equation

$$F(\omega) \equiv \frac{1}{2\pi} \int_{-\infty}^{+\infty} f(t) e^{\pm i\omega t} dt$$

Then

$$F(t) \equiv \int_{-\infty}^{+\infty} f(\omega) e^{\mp i\omega t} d\omega$$

- The Fourier transform of the Gaussian function

$$f(t) = A \exp\left(-\frac{t^2}{t_0^2}\right)$$

is given by

$$F(\omega) = \frac{At_0}{2\sqrt{\pi}} e^{-\frac{\omega^2 t_0^2}{4}}$$

- In general, if a function has a temporal spread of Δt then its Fourier transform $F(\omega)$ will have a spectral spread $\Delta\omega \approx 1/\Delta t$.

PROBLEMS

- 7.1** Consider a periodic force of the form:

$$\begin{aligned} F(t) &= F_0 \sin \omega t \quad \text{for } 0 < t < T/2 \\ &= 0 \quad \text{for } T/2 < t < T \end{aligned}$$

and

$$F(t + T) = F(t)$$

where

$$\omega = \frac{2\pi}{T}$$

Show that

$$\begin{aligned} F(t) &= \frac{1}{\pi} F_0 + \frac{1}{2} F_0 \sin \omega t \\ &\quad - \frac{2}{\pi} F_0 \left(\frac{1}{3} \cos 2\omega t + \frac{1}{15} \cos 4\omega t + \dots \right) \end{aligned}$$

One obtains a periodic voltage of the above form in a half wave rectifier. What will be the Fourier expansion corresponding to full wave rectification?

- 7.2** In quantum mechanics, the solution of the one dimensional Schrödinger equation for a free particle is given by

$$\Psi(x, t) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{+\infty} a(p) e^{\frac{i}{\hbar} \left(px - \frac{p^2}{2m} t \right)} dp$$

where p is the momentum of the particle of mass m . Show that

$$a(p) = \frac{1}{\sqrt{2\pi\hbar}} \int_{-\infty}^{+\infty} \Psi(x, 0) e^{-\frac{i}{\hbar} px} dx$$

7.3 In continuation of Problem 7.2, if we assume

$$\Psi(x, 0) = \frac{1}{(\pi\sigma^2)^{1/4}} \exp\left[-\frac{x^2}{2\sigma^2}\right] \exp\left[\frac{i}{\hbar} p_0 x\right]$$

then show that

$$a(p) = \left(\frac{\sigma^2}{\pi\hbar^2}\right)^{1/4} \exp\left[-\frac{\sigma^2}{2\hbar^2}(p - p_0)^2\right]$$

Also show that

$$\int_{-\infty}^{+\infty} |\Psi(x, 0)|^2 dx = 1 = \int_{-\infty}^{+\infty} |a(p)|^2 dp$$

Indeed $|\Psi(x, 0)|^2 dx$ represents the probability of finding the particle between x and $x + dx$ and $|a(p)|^2 dp$ represents the probability of finding the momentum between p and $p + dp$ and we would have the uncertainty relation

$$\Delta x \Delta p \sim \hbar$$

7.4 Use Eq. (57) to calculate the Fourier transform of the following functions:

$$(a) f(t) = Ae^{-\alpha|t|}$$

$$(b) f(t) = A e^{-t/2\tau} e^{i\omega_0 t} \quad t > 0$$

$$= 0 \quad t < 0$$

REFERENCES AND SUGGESTED READINGS

1. H.S. Carslaw, *Introduction to the Theory of Fourier Series and Integrals*, Dover Publications, New York, 1950.
2. E.C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Oxford University Press, New York, 1937.
3. A. K. Ghatak, I. C. Goyal and S. J. Chua, *Mathematical Physics*, Macmillan India Ltd, New Delhi, 1995.
4. J. Arsac, *Fourier Transforms and the Theory*, Prentice-Hall, Englewood Cliffs, 1966.
5. E.C. Titchmarsh, *Introduction to the Theory of Fourier Integrals*, Clarendon Press, Oxford, (1959).
6. J. W. Goodman, *Introduction to Fourier Optics*, McGraw-Hill, New York, 1968.
7. A. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978.

Chapter 8

Group Velocity and Pulse Dispersion

In a perfect wave, you cannot say when it starts, so you cannot use it for a timing signal. In order to send a signal you have to change the wave somehow, make a notch in it, make it a little bit fatter or thinner. That means that you have to have more than one frequency in the wave, and it can be shown that the speed at which signals travel is not dependent upon the index alone, but upon the way that the index changes with the frequency.

—Richard Feynman in Feynman Lectures in Physics, Vol. I

Important Milestone

- 1672 Isaac Newton reported to the Royal Society his observations on the dispersion of sunlight as it passed through a prism. From this experiment, Newton concluded that sunlight is composed of light of different colours which are refracted by glass to different extents.

8.1 INTRODUCTION

When we switch a light source on and off, we produce a pulse. This pulse propagates through a medium with what is known as the group velocity, which will be discussed in this chapter. In addition, as the pulse propagates, it undergoes distortion which will also be discussed.* A study of this distortion of optical pulses is a subject of great importance in many areas; in particular, it has very important significance in fiber-optic communication systems which will be briefly discussed in Chapter 24.

8.2 GROUP VELOCITY

Let us consider two plane waves (having the same amplitude A) with slightly different frequencies $\omega + \Delta\omega$ and $\omega - \Delta\omega$ propagating along the $+z$ direction:

$$\Psi_1(z, t) = A \cos[(\omega + \Delta\omega)t - (k + \Delta k)z] \quad (1)$$

$$\Psi_2(z, t) = A \cos[(\omega - \Delta\omega)t - (k - \Delta k)z] \quad (2)$$

where $k + \Delta k$ and $k - \Delta k$ are the wave numbers corresponding to the frequencies $\omega + \Delta\omega$ and $\omega - \Delta\omega$ respectively. The superposition of the two waves will be given by

$$\begin{aligned} \Psi(z, t) &= A \cos[(\omega + \Delta\omega)t - (k + \Delta k)z] + \\ &\quad A \cos[(\omega - \Delta\omega)t - (k - \Delta k)z] \end{aligned}$$

or

$$\Psi(z, t) = 2A \cos(\omega t - kz) \cos[(\Delta\omega)t - (\Delta k)z] \quad (3)$$

In Fig. 8.1(a) we have shown the variation of the rapidly varying $\cos(\omega t - kz)$ term at $t = 0$; the distance between two consecutive peaks is $2\pi/k$. In Fig. 8.1(b) we have shown the variation of the slowly varying envelope term, represented by $\cos[(\Delta\omega)t - (\Delta k)z]$ at $t = 0$; the distance between two consecutive peaks is $2\pi/\Delta k$. In Fig. 8.2(a) and (b) we have plotted $\Psi(z, t)$ at

$$t = 0 \quad \text{and} \quad t = \Delta t$$

Obviously the rapidly varying first term moves with the velocity

$$v_p = \frac{\omega}{k} \quad (4)$$

and the slowly varying envelope [which is represented by the second term in Eq. (3)] moves with velocity

$$v_g = \frac{\Delta\omega}{\Delta k} \quad (5)$$

The quantities v_p and v_g are known as the *phase velocity* and the *group velocity* respectively. The group velocity is a

* This chapter assumes a knowledge of waves which will be discussed in the next chapter. May be, the reader would like to go through chapter 9 first before going through this chapter.

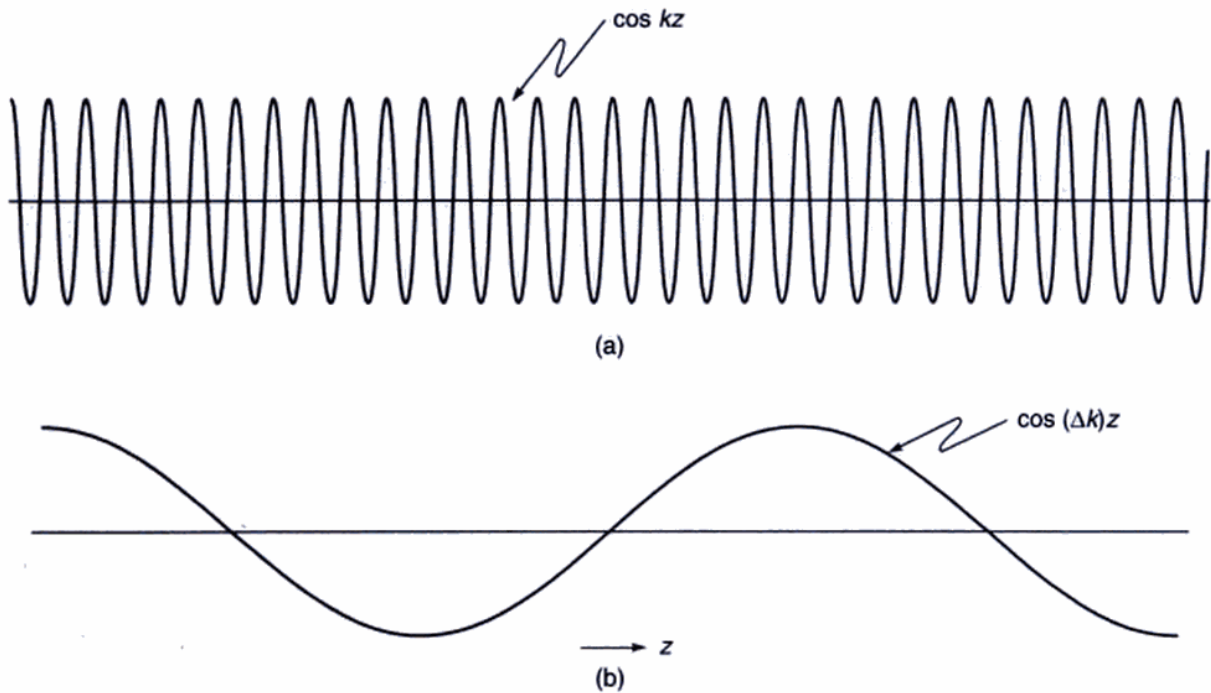


Fig. 8.1 (a) Variation of the rapidly varying $\cos(\omega t - kz)$ term at $t = 0$; the distance between two consecutive peaks is $2\pi/k$. (b) Variation of the slowly varying envelope term, represented by $\cos[(\Delta\omega)t - (\Delta k)z]$, at $t = 0$. The distance between two consecutive peaks is $2\pi/\Delta k$.

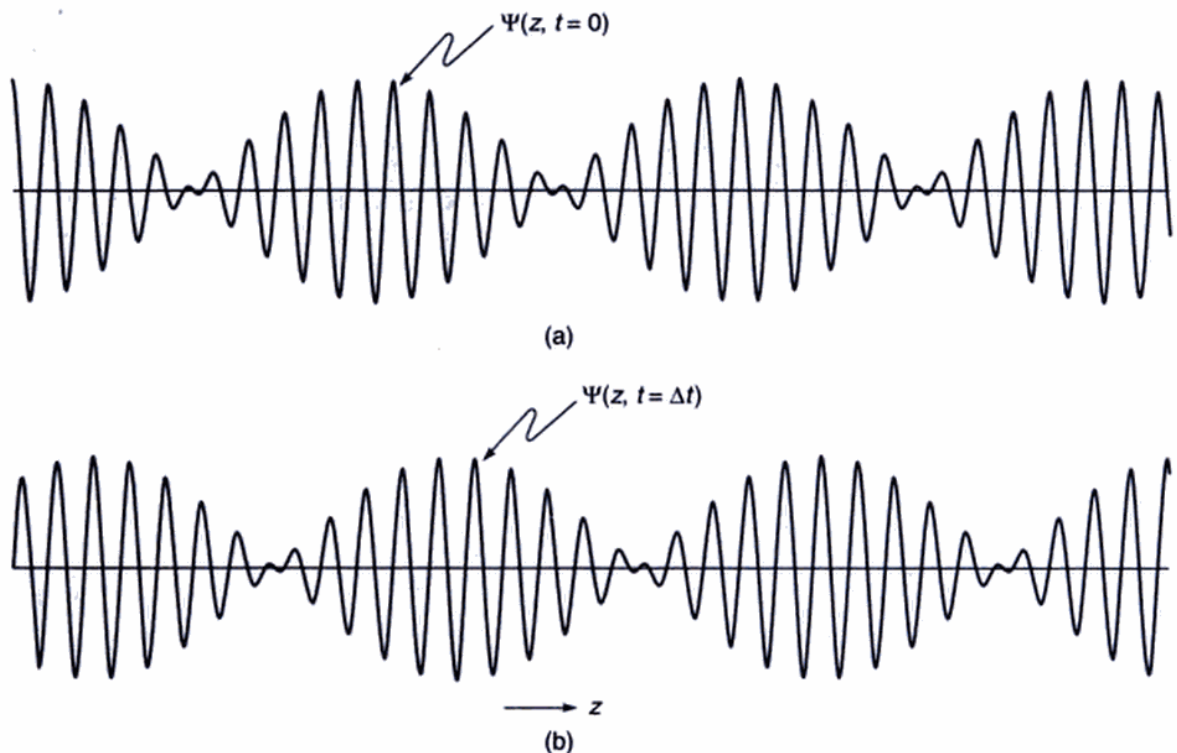


Fig. 8.2 (a) and (b) show the variation of $\Psi(z, t)$ at $t = 0$ and at $t = \Delta t$; the envelope moves with the group velocity $\Delta\omega/\Delta k$.

concept of great importance; indeed in the next section we will rigorously show that a temporal pulse travels with the group velocity given by

$$v_g = \frac{1}{dk/d\omega} \quad (6)$$

Now, in a medium characterized by the refractive index variation $n(\omega)$

$$k(\omega) = \frac{\omega}{c} n(\omega) \quad (7)$$

Thus

$$\frac{1}{v_g} = \frac{dk}{d\omega} = \frac{1}{c} \left[n(\omega) + \omega \frac{dn}{d\omega} \right] \quad (8)$$

In free space $n(\omega) = 1$ at all frequencies; hence

$$v_g = v_p = c \quad (9)$$

Returning to Eq. (8), we may mention that it is customary to express in terms of the free space wavelength λ_0 which is related to ω through the following equation.

$$\omega = \frac{2\pi c}{\lambda_0} \quad (10)$$

Thus

$$\frac{dn}{d\omega} = \frac{dn}{d\lambda_0} \frac{d\lambda_0}{d\omega} = -\frac{\lambda_0^2}{2\pi c} \frac{dn}{d\lambda_0} \quad (11)$$

or,

$$\frac{1}{v_g} = \frac{1}{c} \left[n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0} \right] \quad (12)$$

The group index n_g is defined as

$$n_g = \frac{c}{v_g} = n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0} \quad (13)$$

In Table 8.1 we have tabulated $n(\lambda_0)$, $dn/d\lambda_0$ and $n_g(\lambda_0)$ for pure silica as a function of the free space wavelength λ_0 . In Fig. 8.3 we have plotted (for pure silica) the wavelength variations of the group velocity v_g —we may notice that the group velocity attains a maximum value at $\lambda_0 = 1.27 \mu\text{m}$. As we will show later in this chapter (and also in Chapter 24), this wavelength is of great significance in optical communication systems.

Example 8.1 For pure silica the refractive index variation in the wavelength domain $0.5 \mu\text{m} < \lambda_0 < 1.6 \mu\text{m}$ can be assumed to be given by the following approximate empirical formula:

$$n(\lambda_0) \approx C_0 - a\lambda_0^2 + \frac{a}{\lambda_0^2} \quad (14)$$

where $C_0 = 1.451$, $a \approx 0.003$ and λ_0 is measured in μm . [A more accurate expression for $n(\lambda_0)$ is given in Problem 8.6]. Simple algebra shows

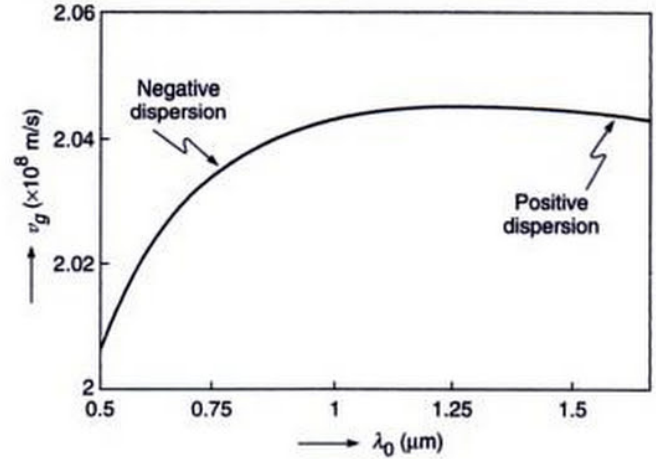


Fig. 8.3 Variation of the group velocity v_g with wavelength for pure silica

$$n_g(\lambda_0) = C_0 + a\lambda_0^2 + \frac{3a}{\lambda_0^2} \quad (15)$$

Thus at $\lambda_0 = 1 \mu\text{m}$,

$$n(\lambda_0) = 1.451$$

and

$$n_g(\lambda_0) = 1.463$$

indicating that the difference between group and phase velocities is about 0.8%. More accurate values of $n(\lambda_0)$ and $n_g(\lambda_0)$ (as obtained by using the expression given in Problem 8.6) are given in Table 8.1.

Using Table 8.1 we find that in pure silica, for

$$\lambda_0 = 0.80 \mu\text{m}, v_g = c/n_g = 2.0444 \times 10^8 \text{ m/s}$$

and for

$$\lambda_0 = 0.85 \mu\text{m}, v_g = c/n_g = 2.0464 \times 10^8 \text{ m/s}$$

implying that (for $\lambda_0 < 1.27 \mu\text{m}$) higher wavelength components travel faster; similarly for $\lambda_0 > 1.27 \mu\text{m}$ lower wavelength components travel faster. Now, every source of light would have a certain wavelength spread, which is usually referred to as the spectral width of the source. Thus a white light source (like coming from the sun) would have a spectral width of about 3000 \AA ; on the other hand, a light emitting diode (usually abbreviated as LED) would have a spectral width of about 25 nm and a typical laser diode (usually abbreviated as LD) operating around $1.3 \mu\text{m}$ would have a spectral width of about 2 nm ; this spectral width is usually denoted by $\Delta\lambda_0$. Since each wavelength component (of a pulse) will travel with a slightly different group velocity it will, in general, result in the broadening of the pulse. In order to calculate this broadening, we note that the time

Table 8.1 Values of n , n_g and D_m for pure silica*

λ_0 (μm)	$n(\lambda_0)$	$\frac{dn}{d\lambda_0}$ (μm^{-1})	$n_g(\lambda_0)$	$\frac{d^2n}{d\lambda_0^2}$ (μm^{-2})	D_m (ps/nm.km)
0.70	1.45561	-0.02276	1.47154	0.0741	-172.9
0.75	1.45456	-0.01958	1.46924	0.0541	-135.3
0.80	1.45364	-0.01725159	1.46744	0.0400	-106.6
0.85	1.45282	-0.01552236	1.46601	0.0297	-84.2
0.90	1.45208	-0.01423535	1.46489	0.0221	-66.4
0.95	1.45139	-0.01327862	1.46401	0.0164	-51.9
1.00	1.45075	-0.01257282	1.46332	0.0120	-40.1
1.05	1.45013	-0.01206070	1.46279	0.0086	-30.1
1.10	1.44954	-0.01170022	1.46241	0.0059	-21.7
1.15	1.44896	-0.01146001	1.46214	0.0037	-14.5
1.20	1.44839	-0.01131637	1.46197	0.0020	-8.14
1.25	1.44783	-0.01125123	1.46189	0.00062	-2.58
1.30	1.44726	-0.01125037	1.46189	-0.00055	2.39
1.35	1.44670	-0.01130300	1.46196	-0.00153	6.87
1.40	1.44613	-0.01140040	1.46209	-0.00235	10.95
1.45	1.44556	-0.01153568	1.46229	-0.00305	14.72
1.50	1.44498	-0.01170333	1.46253	-0.00365	18.23
1.55	1.44439	-0.01189888	1.46283	-0.00416	21.52
1.60	1.44379	-0.01211873	1.46318	-0.00462	24.64

taken by a pulse to traverse a length L of the dispersive medium is given by

$$\tau = \frac{L}{v_g} = \frac{L}{c} \left[n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0} \right] \quad (16)$$

Since the RHS depends on λ_0 , the above equation implies that different wavelengths will travel with different group velocities in propagating through a certain length of the dispersive medium. Thus the pulse broadening will be given by

$$\begin{aligned} \Delta\tau_m &= \frac{d\tau}{d\lambda_0} \Delta\lambda_0 \\ &= -\frac{L\Delta\lambda_0}{\lambda_0 c} \left[\lambda_0^2 \frac{d^2n}{d\lambda_0^2} \right] \end{aligned} \quad (17)$$

The quantity $\Delta\tau_m$ is usually referred as material dispersion because it is due to the material properties of the medium—hence the subscript m . In Eq. (17), the quantity inside the square brackets is dimensionless. Indeed, after propagating through a length L of the dispersive medium, a pulse of temporal width τ_0 will get broadened to τ_f where

$$\tau_f^2 \approx \tau_0^2 + (\Delta\tau_m)^2 \quad (18)$$

In the next section we will explicitly show this for a Gaussian pulse. From Eq. (17) we see that the broadening of the pulse is proportional to the length L traversed in the medium and also to spectral width of the source $\Delta\lambda_0$. We assume

$$\Delta\lambda_0 = 1\text{ nm} = 10^{-9}\text{ m} \quad \text{and} \quad L = 1\text{ km} = 1000\text{ m}$$

and define the dispersion coefficient as

$$D_m = \frac{\Delta\tau_m}{L\Delta\lambda_0} = -\frac{1}{3\lambda_0} \left(\lambda_0^2 \frac{d^2n}{d\lambda_0^2} \right) \times 10^4 \text{ ps/km.nm} \quad (19)$$

where λ_0 is measured in μm and we have assumed $c \approx 3 \times 10^8$ m/s. The quantity D_m is usually referred as the material dispersion coefficient (because it is due to the material properties of the medium) and hence the subscript m on D . A medium is said to be characterized by positive dispersion when D_m is positive and it is said to be characterized by negative dispersion when D_m is negative.

We may mention here that the spectral width of a pulse is usually due to the intrinsic spectral width of the source—which for a typical LED is about 25 nm and for a commercially available laser diode is about 1–2 nm. On the other hand, for a nearly monochromatic source, the intrinsic spectral width could be extremely small and the actual spectral width of a pulse is determined from its finite duration (such

*The numerical values in the Table correspond to the refractive index variation as given in Ref. 2 (see Problem 8.6).

a pulse is often referred to as a Fourier transformed pulse). Thus a 20 ps pulse will have a spectral width

$$\Delta\nu \approx \frac{1}{20 \times 10^{-12}} \approx 5 \times 10^{11} \text{ Hz}$$

implying

$$\Delta\lambda_0 \approx \frac{\lambda_0^2 \Delta\nu}{c} \approx 0.4 \text{ nm}$$

We may see that

$$\frac{d^2n}{d\lambda_0^2} \approx 0$$

around $\lambda_0 \approx 1.27 \mu\text{m}$. Indeed the wavelength $\lambda_0 \approx 1270 \text{ nm}$ is usually referred to as the zero material dispersion wavelength and it is because of low material dispersion, the second and third generation optical communication systems operated around $\lambda_0 \approx 1300 \text{ nm}$; more details will be given in Chapter 24.

Example 8.2 In the I-generation optical communication system, one used LED's with $\lambda_0 = 0.85 \mu\text{m}$ and $\Delta\lambda_0 = 25 \text{ nm}$. Now at $\lambda_0 \approx 0.85 \mu\text{m}$

$$\frac{d^2n}{d\lambda_0^2} \approx 0.030 (\mu\text{m})^{-2}$$

giving

$$D_m \approx -85 \text{ ps/km.nm}$$

the negative sign indicating that higher wavelengths travel faster than lower wavelengths. Thus for $\Delta\lambda_0 \approx 25 \text{ nm}$, the actual broadening of the pulse will be

$$\Delta\tau_m \approx 2.1 \text{ ns/km}$$

implying that the pulse will broaden by 2.1 ns after traversing through 1 km of the silica fiber.

Example 8.3 In the IV-generation optical communication systems, one uses laser diodes with $\lambda_0 = 1.55 \mu\text{m}$ and $\Delta\lambda_0 = 2 \text{ nm}$. Now at $\lambda_0 = 1.55 \mu\text{m}$

$$\frac{d^2n}{d\lambda_0^2} = 0.0042 (\mu\text{m})^{-2}$$

giving

$$D_m = +21.7 \text{ ps/km.nm}$$

the positive sign indicating that higher wavelengths travel slower than lower wavelengths. (Notice from Table 8.1 that for $\lambda_0 \geq 1.27 \mu\text{m}$, n_g increases with λ_0). Thus for $\Delta\lambda_0 \approx 2 \text{ nm}$, the actual broadening of the pulse will be

$$\Delta\tau_m \approx 43 \text{ ps/km}$$

implying that the pulse will broaden by 43 ps after traversing through 1 km of the silica fiber.

8.3 GROUP VELOCITY OF A WAVE PACKET

The displacement corresponding to a one-dimensional plane wave propagating in the +z direction can be written in the form

$$E(z, t) = A e^{i(\omega t - kz)} \quad (20)$$

where A represents the amplitude of the wave and

$$k(\omega) = \frac{\omega}{c} n(\omega) \quad (21)$$

n being the refractive index of the medium. The wave described by Eq. (20) is said to describe a monochromatic wave which propagates with the phase velocity given by

$$v_p = \frac{\omega}{k} = \frac{c}{n} \quad (22)$$

We may mention here that, in general, A may be complex and if we write

$$A = |A| e^{i\phi}$$

then Eq. (20) becomes

$$E = |A| e^{i(\omega t - kz + \phi)}$$

The actual displacement is the real part of E and is, therefore, given by

$$\begin{aligned} \text{Actual electric field} &= \text{Re}(E) \\ &= |A| \cos(\omega t - kz + \phi) \end{aligned} \quad (23)$$

The plane wave represented by Eq. (20) is a practical impossibility because at an arbitrary value of z , the displacement is finite for *all* values of t ; for example,

$$E(z = 0, t) = A e^{+i\omega t}; -\infty < t < \infty \quad (24)$$

which corresponds to a sinusoidal variation for *all* values of time. In practice, the displacement is finite only over a certain domain of time and we have what is known as a wave packet. A wave packet can always be expressed as a superposition of plane waves of different frequencies:

$$E(z, t) = \int_{-\infty}^{+\infty} A(\omega) e^{i[\omega t - kz]} d\omega \quad (25)$$

Obviously

$$E(z = 0, t) = \int_{-\infty}^{+\infty} A(\omega) e^{+i\omega t} d\omega \quad (26)$$

Thus, $E(z = 0, t)$ is the Fourier transform of $A(\omega)$ and using the results of the previous chapter we obtain

$$A(\omega) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} E(z = 0, t) e^{-i\omega t} dt \quad (27)$$

Thus if $E(z = 0, t)$ we know we can determine $E(z, t)$ using the following recipe:

We first determine $A(\omega)$ from Eq. (27) substitute it in Eq. (25) and carry out the resulting integration.

Example 8.4 Gaussian Pulse: As an example, we consider a Gaussian pulse for which we may write

$$E(z = 0, t) = E_0 e^{-\frac{t^2}{\tau_0^2}} e^{+i\omega_0 t} \quad (28)$$

If we substitute Eq. (28) in Eq. (27) we would obtain

$$\begin{aligned} A(\omega) &= \frac{E_0}{2\pi} \int_{-\infty}^{+\infty} e^{-\frac{t^2}{\tau_0^2}} e^{-i(\omega - \omega_0)t} dt \\ &= \frac{E_0 \tau_0}{2\sqrt{\pi}} \exp\left[-\frac{1}{4}(\omega - \omega_0)^2 \tau_0^2\right] \end{aligned} \quad (29)$$

where we have used

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2 + \beta x} dx = \sqrt{\frac{\pi}{\alpha}} e^{\beta^2/4\alpha} \quad (30)$$

(see Appendix A). In general, $A(\omega)$ can be complex and as such one defines the power spectral density

$$S(\omega) = |A(\omega)|^2 \quad (31)$$

For the Gaussian pulse

$$S(\omega) = \frac{E_0^2 \tau_0^2}{4\pi} \exp\left[-\frac{1}{2}(\omega - \omega_0)^2 \tau_0^2\right] \quad (32)$$

In Fig. 8.4 (a) we have plotted the function

$$E_0 e^{-\frac{t^2}{\tau_0^2}} \cos(\omega_0 t)$$

[which is the real part of Eq. (28)] for a 20 fs pulse ($\tau_0 = 20 \times 10^{-15}$ s) corresponding to $\lambda_0 = 1 \mu\text{m}$ ($\omega_0 \approx 6 \pi \times 10^{14}$ Hz); the corresponding spectral density function $S(\omega)$ is plotted in Fig. 8.4(b). As can be seen, $S(\omega)$ is a very sharply peaked function of ω around $\omega = \omega_0$. The full width at half maximum of $S(\omega)$ (usually abbreviated as FWHM) is denoted by $\Delta\omega$; thus at

$$\omega = \omega_0 \pm \frac{1}{2} \Delta\omega$$

$S(\omega)$ attains half of its maximum value; the value of $\Delta\omega$ is obtained from the following equation

$$\frac{1}{2} = \exp\left[-\frac{(\Delta\omega)^2 \tau_0^2}{8}\right]$$

or

$$\text{FWHM} = \Delta\omega = \frac{2\sqrt{2\ln 2}}{\tau_0} \approx \frac{2.36}{\tau_0} \quad (33)$$

Thus the Gaussian pulse of temporal width 20 fs has a frequency spread $\Delta\omega$ given by

$$\Delta\omega \approx 1.18 \times 10^{14} \text{ Hz} \quad (34)$$

Thus

$$\frac{\Delta\omega}{\omega_0} \approx 0.06$$

We may mention here that in order to have clarity in the figure we have chosen a very small value of τ_0 ; usually τ_0 has a much larger value. A larger value of τ_0 will imply a much smaller value of $\Delta\omega$ (resulting in greater monochromaticity of the pulse) and obviously Fig. 8.4(b) will be much more sharply peaked; we will discuss this in greater detail in the chapter on coherence (Chapter 15).

Returning to Eq. (25), we consider the following cases:

8.3.1 Propagation in a Non-Dispersive Medium

For electromagnetic waves, the free space is a non-dispersive medium in which all frequencies propagate with the same velocity c ; thus

$$k(\omega) = \frac{\omega}{c}$$

and Eq. (25) can be written in the form

$$E(z, t) = \int_{-\infty}^{+\infty} A(\omega) e^{-i\frac{\omega}{c}(z-ct)} d\omega \quad (35)$$

The right hand side is a function of $(z - ct)$ and thus any pulse would propagate with velocity c without undergoing any distortion. Thus, for the Gaussian pulse given by Eq. (28).

$$E(z, t) = E_0 e^{-\frac{(z-ct)^2}{c^2 \tau_0^2}} e^{-i\frac{\omega_0}{c}(z-ct)} \quad (36)$$

which represents a distortionless propagation of a Gaussian pulse in a non-dispersive medium*; in Fig. 8.5 we have shown the distortionless propagation of a 20 fs pulse.

*Whereas Eq. (36) follows directly from Eq. (35), it is left as an exercise to the reader to show that if we substitute for $A(\omega)$ from Eq. (29) in Eq. (35), we would readily get Eq. (36).

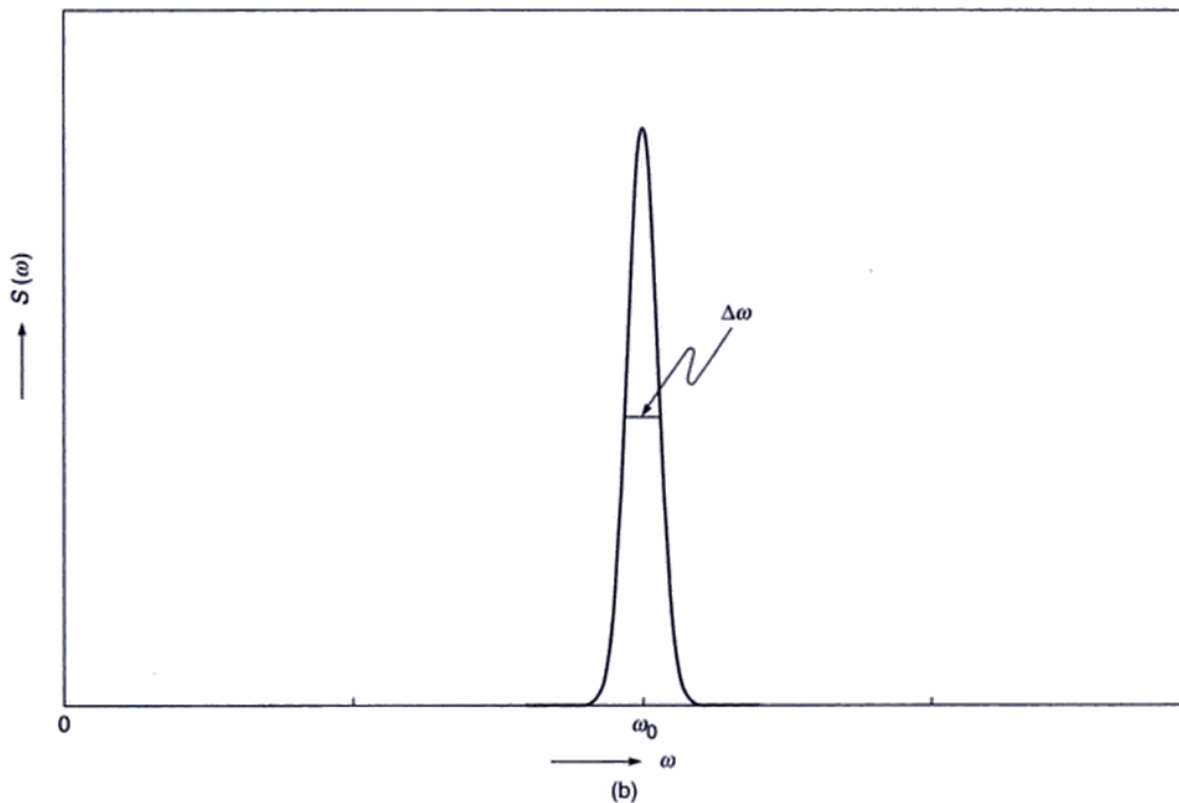
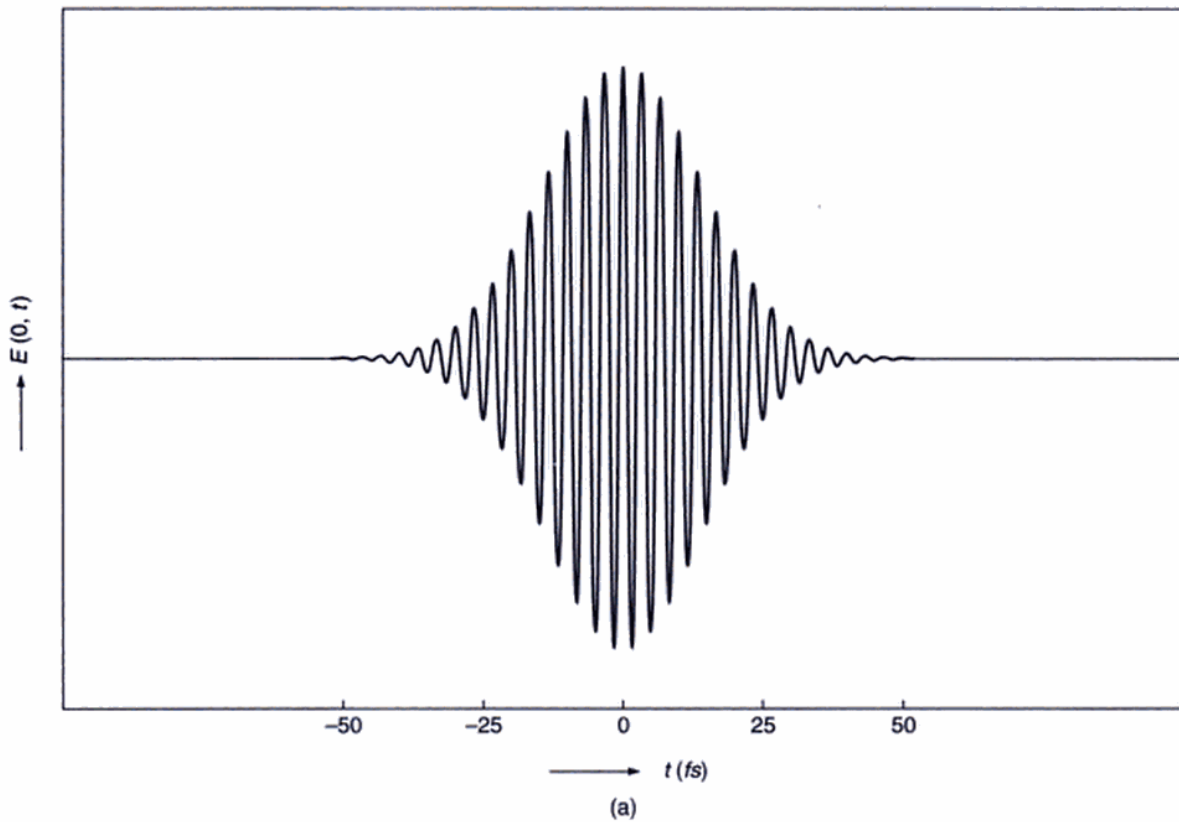


Fig. 8.4 (a) A 20 fs ($= 20 \times 10^{-15}$ s) Gaussian pulse corresponding to $\lambda_0 = 1 \mu\text{m}$; (b) the corresponding frequency spectrum which is usually a very sharply peaked function around $\omega = \omega_0$.

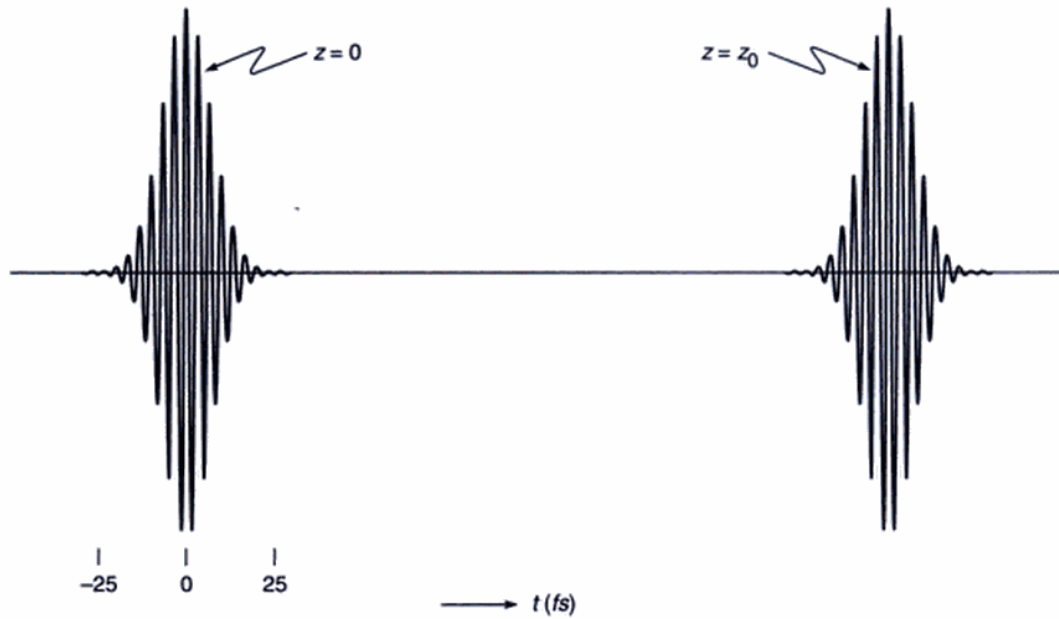


Fig. 8.5 Distortionless propagation of a Gaussian pulse in a non-dispersive medium.

8.3.2 Propagation in a Dispersive Medium

For a wave propagating in a medium characterized by the refractive index variation $n(\omega)$, we will have

$$k(\omega) = \frac{\omega}{c} n(\omega)$$

Now, in most problems, $A(\omega)$ is a very sharply peaked function [see, e.g., Fig. 8.4(b)] so that we may write

$$E(z, t) \approx \int_{\omega_0 - \Delta\omega}^{\omega_0 + \Delta\omega} A(\omega) e^{i[\omega t - k(\omega)z]} d\omega \quad (37)$$

because for $\omega > \omega_0 + \Delta\omega$ and for $\omega < \omega_0 - \Delta\omega$, the function $A(\omega)$ is negligibly small. In this tiny domain of integration, we may make a Taylor series expansion of $k(\omega)$

$$k(\omega) = k(\omega_0) + (\omega - \omega_0) \left. \frac{dk}{d\omega} \right|_{\omega=\omega_0} + \frac{1}{2} (\omega - \omega_0)^2 \left. \frac{d^2k}{d\omega^2} \right|_{\omega=\omega_0} + \dots \quad (38)$$

or

$$k(\omega) = k_0 + \frac{1}{v_g} (\omega - \omega_0) + \frac{1}{2} (\omega - \omega_0)^2 \gamma \quad (39)$$

where

$$k_0 \equiv k(\omega_0) \quad (40)$$

$$\frac{1}{v_g} \equiv \left. \frac{dk}{d\omega} \right|_{\omega=\omega_0} \quad (41)$$

and

$$\gamma \equiv \left. \frac{d^2k}{d\omega^2} \right|_{\omega=\omega_0} \quad (42)$$

We may mention here that we have now defined v_g through Eq. (41)—we will show below that the envelope of the pulse moves with velocity v_g which is the group velocity. Now, if we retain only the first two terms on the RHS of Eq. (39), then Eq. (37) would give us

$$E(z, t) \approx \int_{-\infty}^{+\infty} A(\omega) e^{-i \left[k_0 z + \frac{\omega - \omega_0}{v_g} z - \omega t \right]} d\omega \quad (43)$$

where we have replaced the limits from $-\infty$ to $+\infty$ because, in any case, the contribution from the region $|\omega - \omega_0| > \Delta\omega$ is going to be extremely small. Writing

$$\omega t = (\omega - \omega_0)t + \omega_0 t \quad (44)$$

Eq. (43) can be rewritten in the form

$$E(z, t) \approx \underbrace{e^{i(\omega_0 t - k_0 z)}}_{\text{Phase Term}} \underbrace{\int_{-\infty}^{+\infty} A(\Omega) e^{-\frac{i\Omega}{v_g}(z - v_g t)} d\Omega}_{\text{Envelope Term}} \quad (45)$$

where

$$\Omega \equiv \omega - \omega_0 \quad (46)$$

We see that in the envelope term, z and t do not appear independently but only as $z - v_g t$; thus, the envelope of the pulse moves undistorted with the group velocity

$$v_g = \frac{1}{(dk/d\omega)_{\omega_0}} \quad (47)$$

Thus if we neglect γ [and other higher order terms in Eq. (39)], the pulse moves undistorted with group velocity v_g .

Next, if we take into account all the three terms in Eq. (39), we would obtain

$E(z, t) \approx$

$$\underbrace{e^{i(\omega_0 t - k_0 z)}}_{\text{Phase Term}} \int_{-\infty}^{+\infty} \underbrace{A(\Omega) \exp \left[i\Omega \left(t - \frac{z}{v_g} \right) - \frac{i}{2} \Omega^2 \gamma z \right]}_{\text{Envelope Term}} d\Omega$$

For the Gaussian pulse [see Eq. (28)], $A(\omega)$ is given by Eq. (29); if we now substitute $A(\omega)$ in the above equation and use Eq. (30) to carry out the integration, we would readily obtain

$$E(z, t) = \frac{E_0}{\sqrt{1+ip}} e^{i(\omega_0 t - k_0 z)} \exp \left[-\frac{\left(t - \frac{z}{v_g} \right)^2}{\tau_0^2 (1+ip)} \right] \quad (49)$$

where

$$p \equiv \frac{2\gamma z}{\tau_0^2} \quad (50)$$

The corresponding intensity distribution would be given by

$$I(z, t) = \frac{I_0}{\tau(z)/\tau_0} \exp \left[-\frac{2 \left(t - \frac{z}{v_g} \right)^2}{\tau^2(z)} \right] \quad (51)$$

where

$$\tau^2(z) \equiv \tau_0^2 (1 + p^2) \quad (52)$$

In Fig. 8.6 we have plotted the time variation of the intensity at different values of z . From Eq. (52) we find that as the pulse propagates it undergoes temporal broadening. We define the pulse broadening $\Delta\tau$ as

$$\begin{aligned} \Delta\tau &= \sqrt{\tau^2(z) - \tau_0^2} \\ &= |p| \tau_0 = \frac{2|\gamma|z}{\tau_0} \end{aligned} \quad (53)$$

Now

$$\begin{aligned} \gamma &= \frac{d^2 k}{d\omega^2} = \frac{d}{d\omega} \left[\frac{1}{c} \left(n - \lambda_0 \frac{dn}{d\lambda_0} \right) \right] \\ &= \frac{1}{c} \frac{d}{d\lambda_0} \left[n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0} \right] \frac{d\lambda_0}{d\omega} \\ &= \frac{\lambda_0}{2\pi c^2} \left[\lambda_0^2 \frac{d^2 n}{d\lambda_0^2} \right] \end{aligned} \quad (54)$$

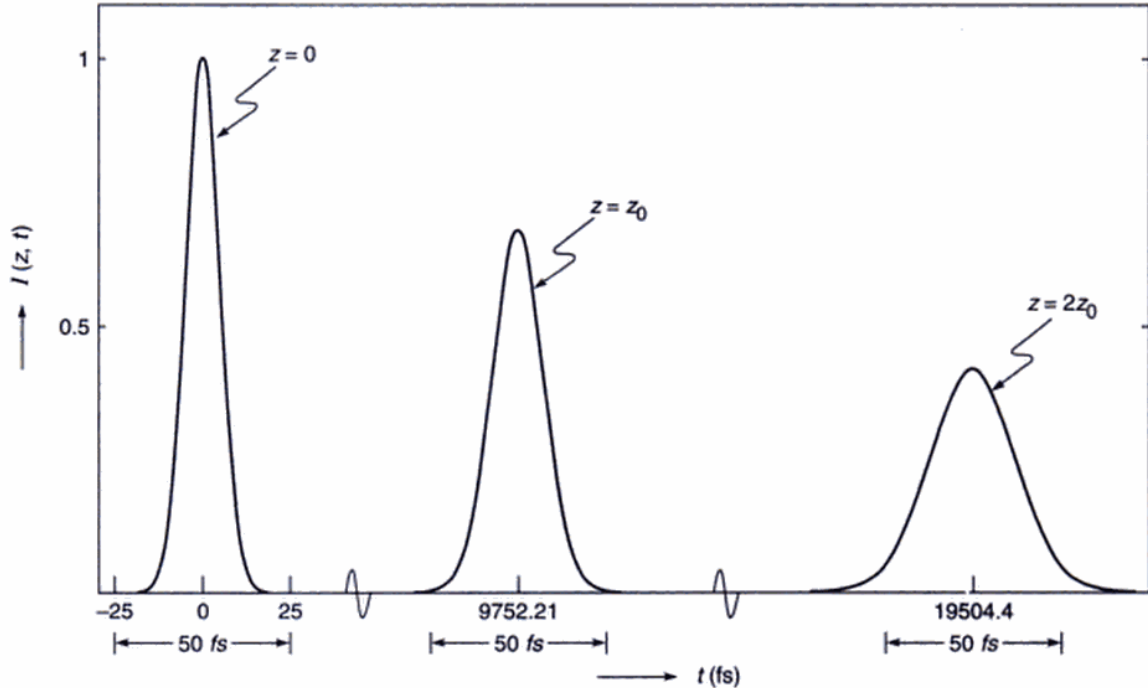


Fig. 8.6 The time variation of the intensity at different values of z ; notice the temporal broadening of the pulse.

where the quantity inside the square brackets is dimensionless. Further, since the spectral width of the Gaussian pulse is given by [see Eq. (33)]

$$\Delta\omega \approx \frac{2}{\tau_0} \quad (55)$$

we may write

$$\frac{1}{\tau_0} \approx \frac{1}{2} \Delta\omega \approx \frac{1}{2} \frac{2\pi c}{\lambda_0^2} |\Delta\lambda_0| \quad (56)$$

Substituting for τ_0 from Eq. (56) and for γ from Eq. (54) in Eq. (53) we get

$$\Delta\tau = \frac{z}{\lambda_0 c} \left| \lambda_0^2 \frac{d^2 n}{d\lambda_0^2} \right| \Delta\lambda_0 \quad (57)$$

which is identical to the result obtained in the earlier section [see Eq. (17)].

Example 8.5 As an example, we assume $\lambda_0 = 1.55 \mu\text{m}$. For pure silica, at this wavelength (see Table 8.1)

$$\frac{d^2 n}{d\lambda_0^2} \approx -0.004165 (\mu\text{m})^{-2}$$

Thus

$$\begin{aligned} \gamma &\approx -\frac{1.55 \times 10^{-6}}{2\pi \times 9 \times 10^{16}} [1.55 \times 1.55 \times 0.004165] \\ &\approx -2.743 \times 10^{-26} \text{ m}^{-1} \text{ s}^2 \end{aligned}$$

For a 100 ps pulse propagating through a 2 km long fiber

$$\Delta\tau \approx \frac{2 \times 2.743 \times 10^{-26} \times 2 \times 10^3}{(10^{-10})} \approx 1.1 \text{ ps}$$

On the other hand, for a 10 fs pulse, at $z = 2 z_0 = 4 \text{ mm}$ we will have

$$\Delta\tau \approx 22 \text{ fs}$$

implying

$$\tau_f \approx [\tau_0^2 + (\Delta\tau)^2]^{1/2} \approx 25 \text{ ps}$$

showing that a 10 fs pulse doubles its temporal width after propagating through a very small distance (see Figs. 8.7 and 8.8).

8.3.3 The Chirping of the Dispersed Pulse

If we carry out simple manipulations, Eq. (49) can be written in the form:

$$E(z, t) = \frac{E_0}{[\tau(z)/\tau_0]^{1/2}} \exp \left[-\frac{\left(t - \frac{z}{v_g} \right)^2}{\tau^2(z)} \right] \times \exp [i(\Phi(z, t) - k_0 z)] \quad (58)$$

where the phase term is given by

$$\Phi(z, t) = \omega_0 t + \kappa \left(t - \frac{z}{v_g} \right)^2 - \frac{1}{2} \tan^{-1} p \quad (59)$$

and

$$\kappa(z) = \frac{p}{\tau_0^2 (1 + p^2)} \quad (60)$$

Equation (59) represents the phase term and the instantaneous frequency is given by

$$\omega(t) = \frac{\partial \Phi}{\partial t} = \omega_0 + 2\kappa \left(t - \frac{z}{v_g} \right) \quad (61)$$

showing that $\omega(t)$ changes within the pulse. The frequency chirp is therefore given by

$$\Delta\omega = \omega(t) - \omega_0 = 2\kappa \left(t - \frac{z}{v_g} \right) \quad (62)$$

Example 8.6 In continuation of Example 8.5, we assume $\lambda_0 = 1.55 \mu\text{m}$ and consider the chirping produced in a 100 ps pulse propagating in pure silica at $z = 2 \text{ km}$. Now

$$\begin{aligned} p &= \frac{2\gamma z}{\tau_0^2} = -\frac{2 \times 2.743 \times 10^{-26} \times 2 \times 10^3}{(100 \times 10^{-12})^2} \\ &\approx -0.011 \end{aligned}$$

At

$$t - \frac{z}{v_g} = -50 \text{ ps}$$

(i.e., at the front end of the pulse)

$$\begin{aligned} \Delta\omega &= \frac{2p}{\tau_0^2 (1 + p^2)} (-50 \times 10^{-12}) \\ &\approx +\frac{2 \times 0.011 \times 50 \times 10^{-12}}{(100 \times 10^{-12})^2} \\ &= +1.1 \times 10^8 \text{ Hz} \end{aligned}$$

Thus at the leading edge of the pulse, the frequencies are slightly higher which is usually referred as 'blue shifted'. Notice

$$\frac{\Delta\omega}{\omega_0} \approx 9 \times 10^{-8}$$

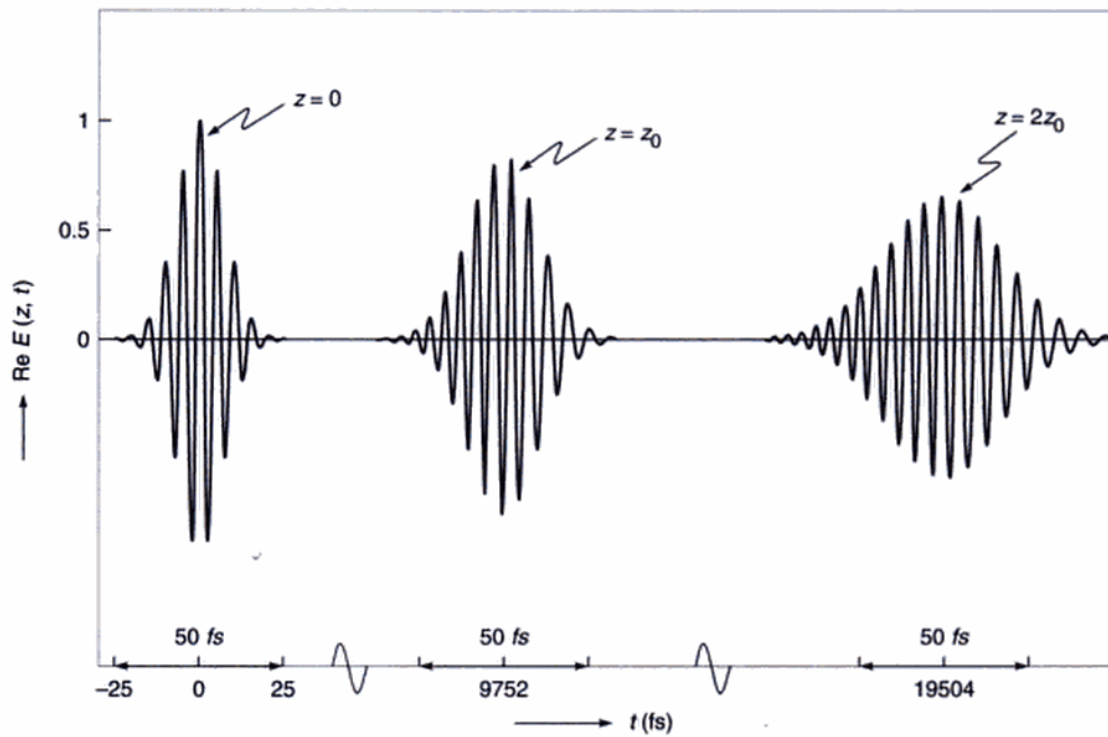


Fig. 8.7 The temporal broadening of a 10 fs unchirped Gaussian pulse ($\lambda_0 = 1.55 \mu\text{m}$) propagating through silica. Notice that since dispersion is positive, the pulse gets down chirped.

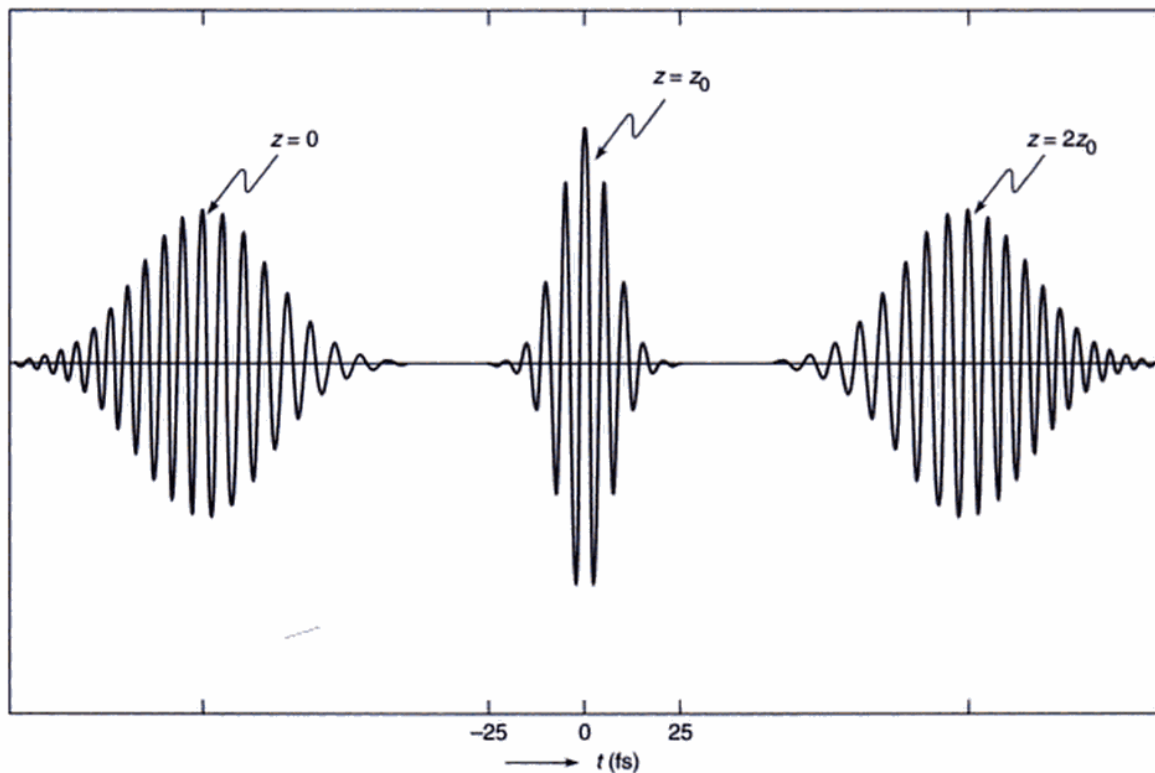


Fig. 8.8 If a down-chirped pulse is passed through a medium characterized by negative dispersion, it will get compressed until it becomes unchirped and then it will broaden again with opposite chirp.

At

$$t = \frac{z}{v_g}, \quad \Delta\omega = 0$$

and at

$$t - \frac{z}{v_g} = +50 \text{ ps}$$

(i.e., at the trailing edge of the pulse)

$$\Delta\omega \approx -1.1 \times 10^8 \text{ Hz}$$

Thus, at the trailing edge of the pulse, the frequencies are slightly lower which is usually referred as 'red-shifted'.

From Example 8.6, we can conclude the following:

For positive dispersion (i.e., negative value of γ), p and κ will also be negative implying that the instantaneous frequency (within the pulse) decreases with time (we are of course assuming $z > 0$); this is known as a **down-chirped pulse** in which the **leading edge** of the pulse ($t < z/v_g$) is **blue-shifted** (i.e., it has frequency higher than ω_0) and the **trailing edge** of the pulse ($t > z/v_g$) is **red-shifted** (i.e., it has frequency lower than ω_0).

This is shown in Fig. 8.7 where at $t = 0$ we have an unchirped pulse. As the pulse propagates further, it will get further broadened and also get further down chirped.

From Eq. (61) it can be readily seen that at negative values of z , p (and therefore κ) will be positive and the **leading edge** of the pulse ($t < z/v_g$) will be **red-shifted** (i.e., it will have frequency lower than ω_0) and the **trailing edge** of the pulse ($t > z/v_g$) will be **blue-shifted** (i.e., it will have frequency higher than ω_0).

This implies that we will have an up-chirped pulse. Thus if an up-chirped pulse is passed through a medium characterized by positive dispersion, it will get compressed until it becomes unchirped and then it will broaden again with opposite chirp.

Similarly we can discuss the case of negative dispersion (implying a positive value of γ). If a down-chirped pulse is passed through a medium characterized by negative dispersion, it will get compressed until it becomes unchirped and then it will broaden again with opposite chirp (see Fig. 8.8).

8.4 SELF PHASE MODULATION

It may be mentioned that as a pulse propagates through a dispersive medium, the frequency spectrum remains the same—i.e., no new frequencies are generated. Different frequencies superpose with different phases to distort the temporal shape of the pulse (see Problem 8.10). New

frequencies are generated when the medium is non-linear — we will briefly discuss this here.

The refractive index of any material is a constant only for small intensities of the propagating laser beam. If the intensities are large, the refractive index variation is approximately given by

$$n \approx n_0 + n_2 I \quad (63)$$

where n_2 is a constant and I represents the intensity of the beam. For example, for fused silica, $n_0 \approx 1.47$ and $n_2 \approx 3.2 \times 10^{-20} \text{ m}^2/\text{W}$. Further, if the effective area of the light beam is A_{eff} , then the intensity is given by

$$I = \frac{P}{A_{\text{eff}}} \quad (64)$$

where P is the power associated with the light beam. Now in a single mode fiber, the spot size w_0 of the beam is about $5 \mu\text{m}$ (see Examples 24.8 and 24.9). Thus the effective cross-sectional area of the beam, $A_{\text{eff}} \approx \pi w_0^2 \approx 50 \mu\text{m}^2$. For a 5 mW laser beam propagating through such a fiber, the resultant intensity is given by

$$I = \frac{P}{A_{\text{eff}}} = \frac{5 \times 10^{-3} \text{ W}}{50 \times 10^{-12} \text{ m}^2} = 10^8 \text{ W/m}^2 \quad (65)$$

Thus the change in refractive index is given by

$$\Delta n = n_2 I \approx 3.2 \times 10^{-12} \quad (66)$$

Although this is very small, but when the beam propagates over an optical fiber over long distances (a few hundred to a few thousand kilometers), the accumulated nonlinear effects can be significant. That is the great advantage of the optical fiber - the beam remains confined to a very small area for long distances!

We consider a laser pulse (of frequency ω_0) propagating through an optical fiber; the effective propagation constant is given by

$$\begin{aligned} k &= \frac{\omega_0}{c} [n_0 + n_2 I] \\ &= \frac{\omega_0}{c} \left[n_0 + n_2 \frac{P(t)}{A_{\text{eff}}} \right] \end{aligned} \quad (67)$$

Thus, for such a propagating beam, the phase term is approximately given by

$$e^{+i(\omega_0 t - kz)} = \exp \left[+i \left\{ \omega_0 t - \frac{\omega_0}{c} \left(n_0 + n_2 \frac{P(t)}{A_{\text{eff}}} \right) z \right\} \right] = e^{+i\Phi}$$

where the phase Φ is defined as

$$\Phi(z, t) \equiv \omega_0 t - \frac{\omega_0}{c} \left(n_0 + n_2 \frac{P(t)}{A_{\text{eff}}} \right) z \quad (68)$$

*Values adapted from Ref. 2.

We can define an instantaneous frequency as [cf. Eq. (61)]:

$$\omega(t) \equiv \frac{\partial \Phi}{\partial t} = \omega_0 - g \frac{dP(t)}{dt} z$$

where

$$g = \frac{n_2 \omega_0}{c A_{eff}} = \frac{2\pi n_2}{\lambda_0 A_{eff}} \quad (69)$$

For $A_{eff} \approx 50 \mu\text{m}^2$, $\lambda_0 = 1.55 \mu\text{m}$ and $n_2 = 3.2 \times 10^{-20} \text{ m}^2/\text{W}$, $g \approx 2.6 \times 10^{-3} \text{ W}^{-1} \text{ m}^{-1}$.

Now, for a Gaussian pulse propagating with group velocity v_g [see Eq. (51)]:

$$P(z, t) = P_0 \exp \left[-\frac{2 \left(t - \frac{z}{v_g} \right)^2}{\tau_0^2} \right]$$

where we have neglected dispersion [i.e., $p = 0$ in Eqs. (49) and (52)]. Thus

$$\omega(t) = \omega_0 \left[1 + \frac{2gz}{\omega_0 \tau_0^2} P_0 \left(t - \frac{z}{v_g} \right) \exp \left[-\frac{2 \left(t - \frac{z}{v_g} \right)^2}{\tau_0^2} \right] \right]$$

For $\lambda_0 = 1.55 \text{ nm}$

$$\omega_0 = \frac{2\pi c}{\lambda_0} = \frac{2\pi \times 3 \times 10^8}{1.55 \times 10^{-6}} \approx 1.22 \times 10^{15} \text{ s}^{-1}$$

Further, for $P_0 = 15 \text{ mW}$, $\tau_0 = 20 \text{ fs}$ and $z = 200 \text{ km}$

$$\begin{aligned} & \frac{2gzP_0}{\omega_0 \tau_0^2} \left(t - \frac{z}{v_g} \right) \\ &= \frac{2 \times 2.6 \times 10^{-3} \times 2 \times 10^5 \times 15 \times 10^{-3}}{1.22 \times 10^{15} \times (20 \times 10^{-15})^2} \left(t - \frac{z}{v_g} \right) \\ &\approx 3.2 \times 10^{13} \left(t - \frac{z}{v_g} \right) \\ &\approx +0.64 \quad \text{for } t - \frac{z}{v_g} \approx 20 \text{ fs (trailing edge of the pulse)} \\ &\approx -0.64 \quad \text{for } t - \frac{z}{v_g} \approx -20 \text{ fs (trailing end of the pulse)} \end{aligned}$$

Thus the instantaneous frequency within the pulse changes with time leading to chirping of the pulse as shown in Fig. 8.9; this is known as self phase modulation (usually abbreviated as SPM). Note that since the pulse width has not changed, but the pulse is chirped, the frequency content of the pulse has increased. Thus SPM leads to generation of new frequencies. Indeed by passing a pulse through a fiber characterized by very small cross-sectional area (so that the value of g is large) it is possible to generate the entire visible spectrum (see Fig. 8.10).

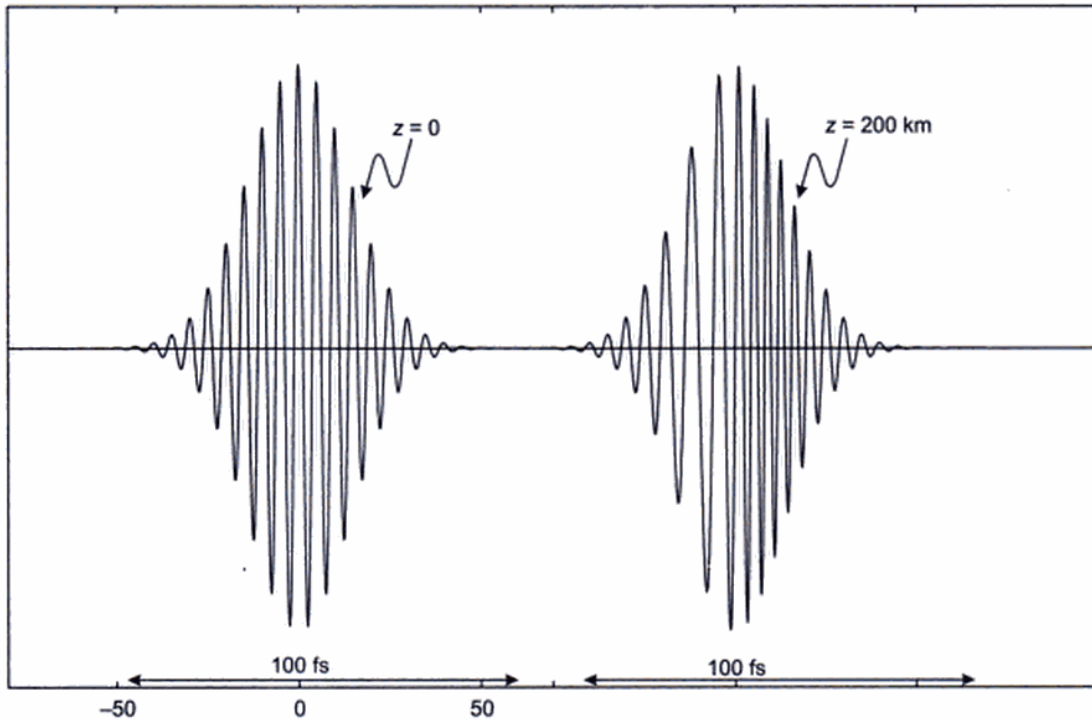


Fig. 8.9 Due to self phase modulation, the instantaneous frequency within the pulse changes with time leading to chirping of the pulse. Calculations correspond to $P_0 = 15 \text{ mW}$, $\lambda_0 = 1550 \text{ nm}$, $\tau_0 = 20 \text{ fs}$, $A_{eff} = 50 \mu\text{m}^2$ and $v_g = 2 \times 10^8 \text{ m/s}$.

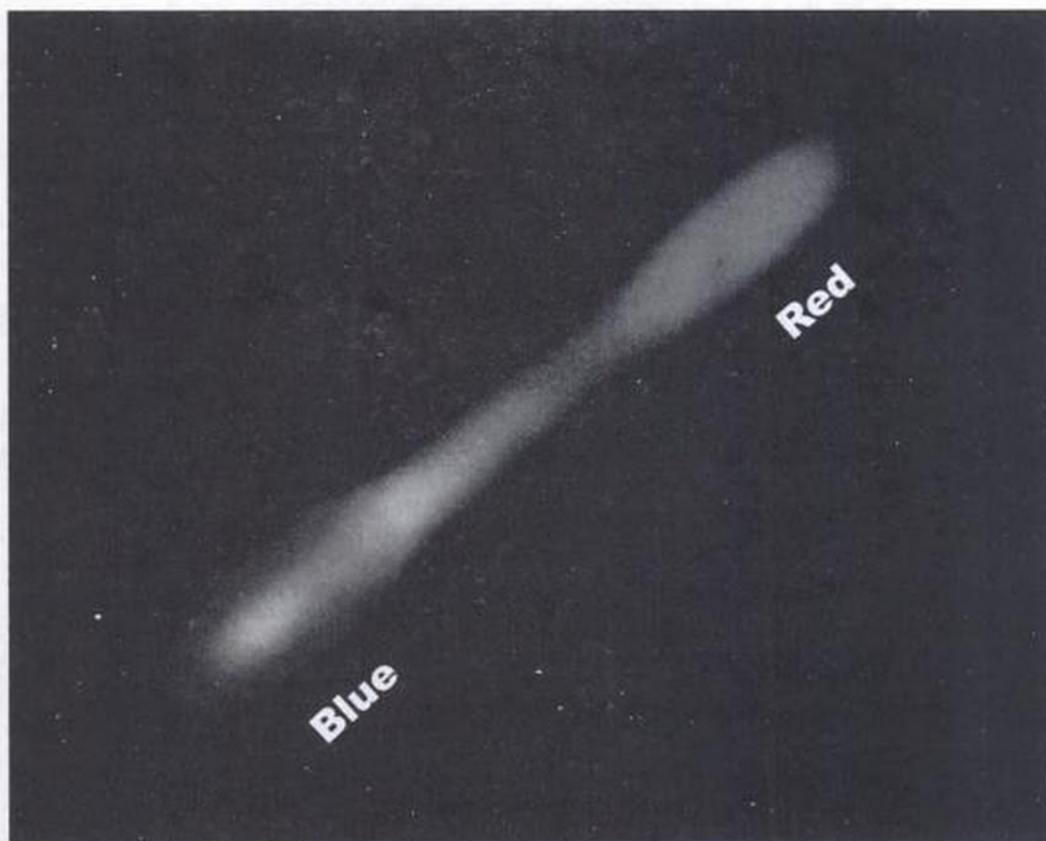


Fig. 8.10 The entire visible spectrum is generated when a short duration pulse propagates through an optical fiber. [Photograph adapted from Optics and Photonics News, October, 1999; the photograph is by J. Ranka, R. Windeler, and A. Stentz].

SUMMARY

- When we switch a light source on and off, we produce a pulse. This pulse propagates through a medium with what is known as the group velocity, which is given by

$$v_g = \frac{1}{dk/d\omega}$$

For a medium characterized by the refractive index variation $n(\omega)$

$$k(\omega) = \frac{\omega}{c} n(\omega),$$

the group velocity is given by

$$\frac{1}{v_g} = \frac{1}{c} \left[n(\lambda_0) - \lambda_0 \frac{dn}{d\lambda_0} \right]$$

where λ_0 is the wavelength in free space and $c \approx (3 \times 10^8 \text{ m/s})$ is the speed of light in free space.

- After traversing through a distance L in a dispersive medium, a pulse will broaden by an amount

$$\Delta t_m = -\frac{L\Delta\lambda_0}{\lambda_0 c} \left[\lambda_0^2 \frac{d^2 n}{d\lambda_0^2} \right]$$

where $\Delta\lambda_0$ is the spectral width of the source; the subscript m denotes that the fact we are considering material dispersion. The dispersion coefficient is given by

$$D_m = \frac{\Delta t_m}{L\Delta\lambda_0} \approx -\frac{1}{3\lambda_0} \left(\lambda_0^2 \frac{d^2 n}{d\lambda_0^2} \right) \times 10^4 \text{ ps/km.nm}$$

where λ_0 is measured in μm and we have assumed $c \approx 3 \times 10^8 \text{ m/s}$. For example, for silica, at $\lambda_0 = 1.55 \mu\text{m}$, $d^2 n/d\lambda_0^2 \approx -0.00416 \mu\text{m}^{-2}$ and $D_m \approx +22$ picoseconds per kilometer (length of the medium) per nanometer (spectral width of the source). On the other

hand, for silica $\frac{d^2 n}{d\lambda_0^2} \approx 0$ around $\lambda_0 \approx 1.27 \mu\text{m}$. Indeed

the wavelength $\lambda_0 \approx 1.27 \mu\text{m}$ is usually referred to as the zero material dispersion wavelength and it is because of low material dispersion, the second and third generation optical communication systems operated around $\lambda_0 \approx 1.3 \mu\text{m}$.

- For a Gaussian pulse

$$E(z=0, t) = E_0 \exp\left(-\frac{t^2}{\tau_0^2}\right) e^{+i\omega_0 t}$$

the temporal width after propagating through a distance z is given by $\tau(z) = \tau_0 \sqrt{1+p^2}$; thus the temporal broadening is given by

$$\Delta\tau = \sqrt{\tau^2(z) - \tau_0^2} = |p| \tau_0$$

where

$$p = \frac{2}{\tau_0^2} \cdot \frac{\lambda_0}{2\pi c^2} \left[\lambda_0^2 \frac{d^2 n}{d\lambda_0^2} \right] \cdot z$$

Thus at $\lambda_0 \approx 1.55 \mu\text{m}$, for a $\tau_0 \approx 100 \text{ ps}$ pulse (propagating in pure silica), $\Delta\tau \approx 0.55 \text{ ps/km}$.

PROBLEMS

- 8.1 Using the empirical formula given by Eq. (14) calculate the phase and group velocities in silica at $\lambda_0 = 0.7 \mu\text{m}$, $0.8 \mu\text{m}$, $1.0 \mu\text{m}$, $1.2 \mu\text{m}$ and $1.4 \mu\text{m}$. Compare with the (more accurate) values given in Table 8.1.
- 8.2 For pure silica we may assume the empirical formula

$$n(\lambda_0) \approx 1.451 - 0.003 \left(\lambda_0^2 - \frac{1}{\lambda_0^2} \right)$$

where λ_0 is measured in μm .

- (a) Calculate the zero dispersion wavelength.
- (b) Calculate the material dispersion at 800 nm in ps/km.nm .

[Ans: $1.32 \mu\text{m}$; -101 ps/km.nm]

- 8.3 Let $n(\lambda_0) = n_0 + A\lambda_0$, where λ_0 is the free space wavelength. Derive expressions for phase and group velocities.

[Ans: $v_g = c/n_0$]

- 8.4 Consider a LED source emitting light of wavelength 850 nm and having a spectral width of 50 nm . Using Table 8.1, calculate the broadening of a pulse propagating in pure silica.

[Ans: 4.2 ns/km]

- 8.5 In 1836 Cauchy gave the following approximate formula to describe the wavelength dependence of refractive index in glass in the visible region of the spectrum:

$$n(\lambda) = A + \frac{B}{\lambda_0^2}$$

Now*

$$\left. \begin{aligned} n(\lambda_1) &= 1.50883 \\ n(\lambda_2) &= 1.51690 \end{aligned} \right\} \text{ for borosilicate glass}$$

$$\left. \begin{aligned} n(\lambda_1) &= 1.45640 \\ n(\lambda_2) &= 1.46318 \end{aligned} \right\} \text{ for vitreous quartz}$$

where $\lambda_1 = 0.6563 \mu\text{m}$ and $\lambda_2 = 0.4861 \mu\text{m}$.

- (a) Calculate the values of A and B .
- (b) Using the Cauchy formula calculate the refractive index at $0.5890 \mu\text{m}$ and $0.3988 \mu\text{m}$ and compare with the following experimental values:

- (i) 1.51124 and 1.52546 for borosilicate glass and
- (ii) 1.45845 and 1.47030 for vitreous quartz.

- 8.6 The refractive index variation for pure silica in the wavelength region $0.5 \mu\text{m} < \lambda_0 < 1.6 \mu\text{m}$ is accurately described by the following empirical formula

$$n(\lambda_0) = C_0 + C_1 \lambda_0^2 + C_2 \lambda_0^4 + \frac{C_3}{(\lambda_0^2 - l)} + \frac{C_4}{(\lambda_0^2 - l)^2} + \frac{C_5}{(\lambda_0^2 - l)^3}$$

where $C_0 = 1.4508554$, $C_1 = -0.0031268$, $C_2 = -0.0000381$, $C_3 = 0.0030270$, $C_4 = -0.0000779$, $C_5 = 0.0000018$, $l = 0.035$ and λ_0 is measured in μm . Calculate and plot $n(\lambda_0)$ and $d^2 n/d\lambda_0^2$ in the wavelength domain $0.5 < \lambda_0 < 1.6 \mu\text{m}$.

- 8.7 (a) For a Gaussian pulse given by

$$E = E_0 e^{-\frac{t^2}{\tau_0^2}} e^{i\omega_0 t}$$

the spectral width is approximately given by

$$\Delta\omega \approx \frac{1}{\tau_0}$$

Assume $\lambda_0 = 8000 \text{ \AA}$.

Calculate $\frac{\Delta\omega}{\omega_0}$ for $\tau_0 = 1 \text{ ns}$ and for $\tau_0 = 1 \text{ ps}$.

- (b) For such a Gaussian pulse, the pulse broadening is given by

$$\Delta\tau = \frac{2z}{\tau_0} |\gamma|$$

where

$$\gamma = \frac{d^2 k}{d\omega^2}$$

Using Table 8.1, calculate $\Delta\tau$ and interpret the result physically.

- 8.8 As a Gaussian pulse propagates, the frequency chirp is given by

$$\Delta\omega = \frac{2p}{\tau_0^2(1+p^2)} \left(t - \frac{z}{v_g} \right)$$

where p is defined in the previous problem. Assume a 100 ps ($= \tau_0$) pulse at $\lambda_0 = 1 \mu\text{m}$. Calculate the frequency chirp

$$\frac{\Delta\omega}{\omega_0}$$

at $t - z/v_g = -100$ ps, -50 ps, $+50$ ps and $+100$ ps. Assume $z = 1$ km and other values from Table 8.1.

8.9 Repeat the previous problem for $\lambda_0 = 1.5 \mu\text{m}$; the values of τ_0 and z remain the same. Discuss the qualitative difference in the results obtained in the various problem.

8.10 The frequency spectrum of $E(0, t)$ is given by the function $A(\omega)$. Show that the frequency spectrum of $E(z, t)$ is simply

$$A(\omega)e^{-ik(\omega)z}$$

implying that no new frequencies are generated—different frequencies superpose with different phases at different values of z .

8.11 The time evolution of a Gaussian pulse in a dispersive medium is given by

$$E(z, t) = \frac{E_0}{\sqrt{1+ip}} e^{i(\omega_0 t - k_0 z)} \exp \left[-\frac{\left(t - \frac{z}{v_g} \right)^2}{\tau_0^2 (1+ip)} \right]$$

where

$$p \equiv \frac{2\gamma z}{\tau_0^2}$$

Calculate explicitly frequency spectrum of $E(0, t)$ and $E(z, t)$ and show that the results agree with that of Problem 8.10.

REFERENCES AND SUGGESTED READINGS

1. R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley Publishing Co., Reading, Massa., 1964.
2. U.C. Paek, G.E. Peterson and A. Carnevale, 'Dispersionless single mode light guides with a index profiles' Bell System Technical Journal, Vol. 60, 583, 1981.
3. A. Ghatak and K Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, Cambridge, 1998, (Reprinted in India by Foundation Books, New Delhi).

Chapter 9

Wave Propagation and the Wave Equation

'If you are dropping pebbles into a pond and do not watch the spreading rings, your occupation should be considered as useless', said the fictional Russian philosopher, Kuzma Prutkoff. And, indeed we can learn much by observing these graceful circles spreading out from the punctured surface of calm water.

—Gamow and Cleveland

9.1 INTRODUCTION

In this chapter we will discuss the phenomenon of waves. A wave is propagation of a disturbance. For example, when we drop a small stone in a calm pool of water, a circular pattern spreads out from the point of impact. The impact of the stone creates a disturbance which propagates outwards. In this propagation, the water molecules do not move outward with the wave; instead they move in nearly circular orbits about an equilibrium position. Once the disturbance has passed a certain region, every drop of water is left at its original position. This fact can easily be verified by placing a small piece of wood on the surface of water. As the wave passes, the piece of wood makes oscillations and once the disturbance has passed, the wood comes back to its original position. Further, with time the circular ripples spread out, i.e., the disturbance (which is confined to a particular region at a given time) produces a similar disturbance at a neighbouring point at a slightly later time with the pattern of disturbance roughly remaining the same. Such propagation of disturbances (without any translation of the medium in the direction of propagation) is termed as a wave. It is also seen that the wave carries energy; in this case the energy is in the form of kinetic energy of water molecules.

We will first consider the simplest example of wave propagation, viz., the propagation of a transverse wave on a string. Consider yourself holding one end of a string, the other end being held tightly by another person so that the string does not sag. If you move the end of the string up and down a few times then a disturbance is created which propagates towards the other end of the string. Thus, if we take a snapshot of the string at $t = 0$ and at a slightly later time Δt , then the snapshots will roughly* look like the ones

shown in Figs 9.1 (a) and (b). The figures show that the disturbances have identical shapes except for the fact that one is displaced from the other by distance $v \Delta t$ where v represents the speed of the disturbance. Such a propagation of a disturbance without its change in form is a characteristic of a wave. The following points may, however, be noted:

- (a) A certain amount of work is done when the wave is generated and as the wave propagates through the string, it carries with it a certain amount of energy which is felt by the person holding the other end of the string.
- (b) The wave is transverse, i.e., the displacement of the particles of the string is at right angles to the direction of propagation of the wave.

Referring back to Figs 9.1 (a) and (b), we note that the shape of the string at the instant Δt is similar to its shape at $t = 0$, except for the fact that the whole disturbance has travelled through a certain distance. If v represents the speed of the wave then this distance is simply $v \Delta t$. Consequently, if the equation describing the rope at $t = 0$ is $y(x)$ then at a later instant t , the equation of the curve would be $y(x - vt)$ which simply implies a shift of the origin by a distance vt . Similarly, for a disturbance propagating in the $-x$ direction, if the equation describing the rope at $t = 0$ is $y(x)$, then at a later instant t the equation of the curve would be $y(x + vt)$.

Example 9.1 Study the propagation of a semicircular pulse in the $+x$ direction whose displacement at $t = 0$ is given by the following equations:

$$\begin{aligned} y(x, t = 0) &= [R^2 - x^2]^{1/2} & |x| \leq R \\ &= 0 & |x| \geq R \end{aligned} \quad (1)$$

*We are assuming here that as the disturbance propagates through the string, there is negligible attenuation and also no change in the shape of the disturbance.

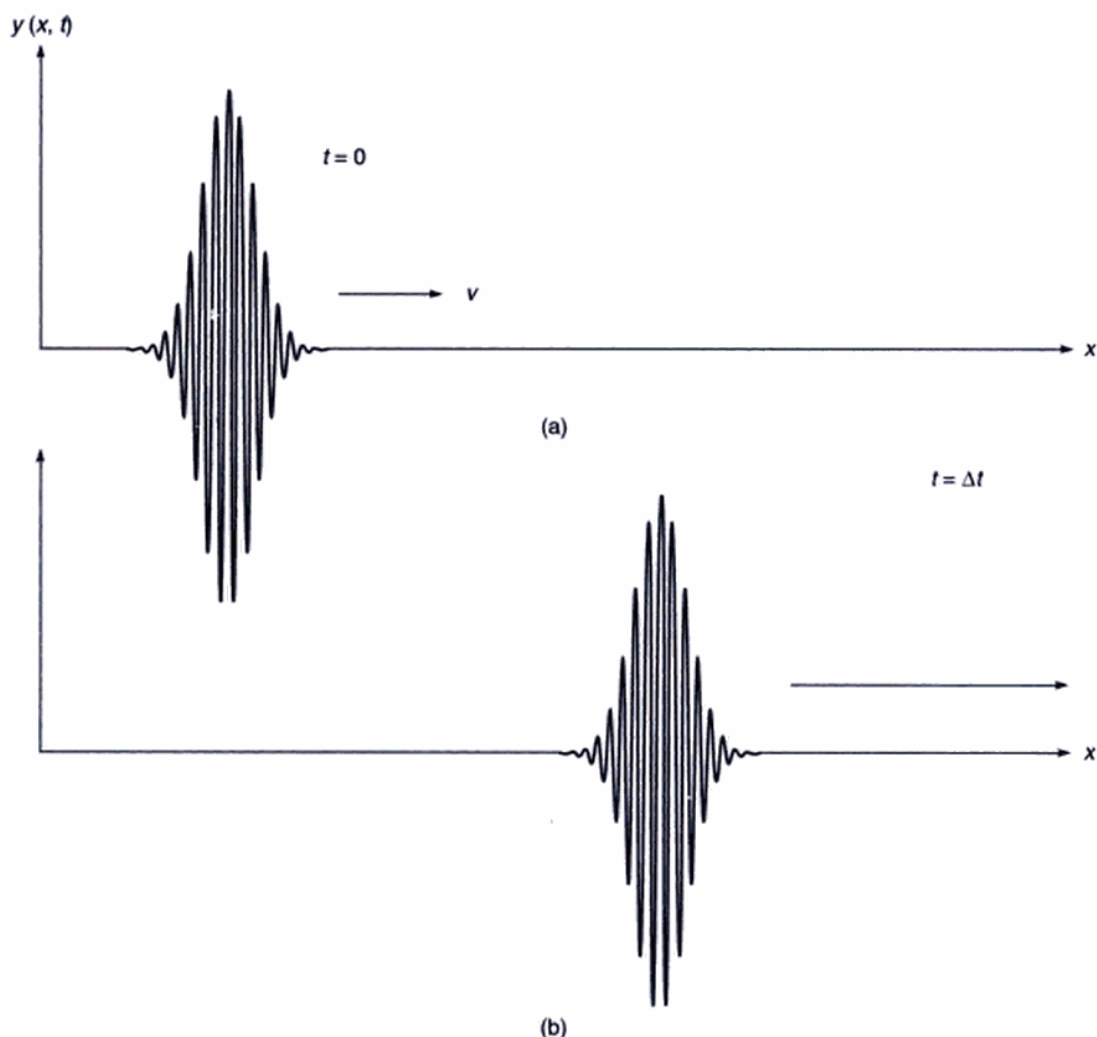


Fig. 9.1 A transverse wave is propagating along the $+x$ -axis on a string; (a) and (b) show the displacements at $t = 0$ and $t = \Delta t$ respectively.

Solution: For a wave propagating in the $+x$ direction the dependence of $y(x, t)$ on x and t should be through the function $(x - vt)$. Consequently,

$$y(x, t) = [R^2 - (x - vt)^2]^{1/2} \quad |x - vt| \leq R$$

$$= 0 \quad |x - vt| \geq R \quad (2)$$

The shape of the pulse at $t = 0$ and at a later time t_0 is shown in Fig. 9.2. Equation (2) immediately follows from the fact that $y(x, t)$ has to be of the form $y(x - vt)$ and at $t = 0$, $y(x, t)$ must be given by Eq. (1).

Example 9.2 Consider a pulse propagating in the minus x -direction with speed v . The shape of the pulse at $t = t_0$ is given by

$$y(x, t = t_0) = \frac{b^2}{a^2 + (x - x_0)^2} \quad (3)$$

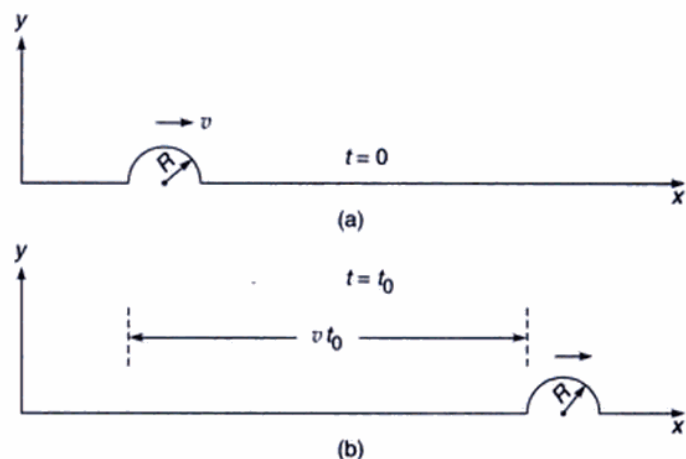


Fig. 9.2 The propagation of a semicircular pulse along the $+x$ -axis; (a) and (b) show the shape of the pulse at $t = 0$ and at a later time t_0 respectively.

(Such a pulse is known as a Lorentzian pulse.) Determine the shape of the pulse at an arbitrary time t .

Solution: The shape of the pulse at $t = t_0$ is shown in Fig. 9.3 (a). The maximum of the displacement occurs at $x = x_0$. Since the pulse is propagating in the $-x$ direction, at a later time t , the maximum will occur at $x_0 - v(t - t_0)$. Consequently, the shape of the pulse at an arbitrary time t would be given by

$$y(x, t) = \frac{b^2}{a^2 + [x - x_0 + v(t - t_0)]^2} \quad (4)$$

Equation (4) could have been written down directly from Eq. (3) by replacing x by $x + v(t - t_0)$.

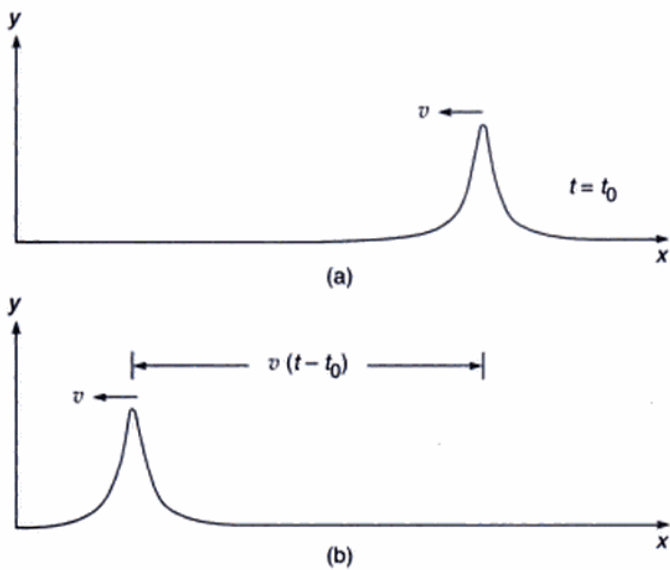


Fig. 9.3 The propagation of a Lorentzian pulse along the minus x -axis; (a) and (b) show the shape of the pulse at $t = t_0$ and at a later instant t respectively.

9.2 SINUSOIDAL WAVES: CONCEPT OF FREQUENCY AND WAVELENGTH

Till now we have been considering the propagation of a pulse which lasts for a finite amount of time. We will now consider a periodic wave in which the displacement $y(x, t)$ has the form

$$y(x, t) = a \cos [k(x \mp vt) + \phi] \quad (5)$$

where the upper and lower signs correspond to waves propagating in the $+x$ and $-x$ directions respectively. Such a displacement is indeed produced in a long stretched string

at the end of which a continuously vibrating tuning fork is placed. The quantity ϕ is known as the phase of the wave (see Chapter 6). We may, without loss of generality, assume $\phi = 0$. Thus for a wave propagating in the $+x$ direction,

$$y(x, t) = a \cos k(x - vt) \quad (6)$$

In Fig. 9.4 we have plotted the dependence of the displacement y on x at $t = 0$ and at $t = \Delta t$. These are given by

$$y(x) = a \cos kx \quad \text{at } t = 0 \quad (7)$$

and

$$y(x) = a \cos k(x - v\Delta t) \quad \text{at } t = \Delta t$$

The two curves are the snapshots of the string at the two instants. It can be seen from the figure that, at a particular instant, any two points separated by a distance

$$\lambda = 2\pi/k \quad (8)$$

have identical displacements. This distance is known as the wavelength. Further, the displaced curve (which corresponds to the instant $t = \Delta t$) can be obtained by displacing the curve corresponding to $t = 0$ by a distance $v\Delta t$; this shows that the wave is propagating in the $+x$ direction with speed v . It can also be seen that the maximum displacement of the particle (from its equilibrium position) is a , which is known as the amplitude of the wave.

In Fig. 9.5 we have plotted the time dependence of the displacement of the points characterised by $x = 0$ and $x = \Delta x$. These are given by

$$\text{and } \left. \begin{aligned} y(t) &= a \cos \omega t & \text{at } x = 0 \\ y(t) &= a \cos (\omega t - k\Delta x) & \text{at } x = \Delta x \end{aligned} \right\} \quad (9)$$

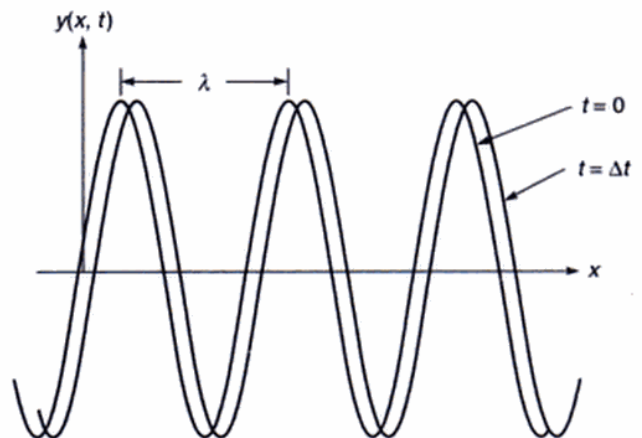


Fig. 9.4 The curves represent the displacement of a string at $t = 0$ and at $t = \Delta t$ respectively when a sinusoidal wave is propagating in the $+x$ -direction.

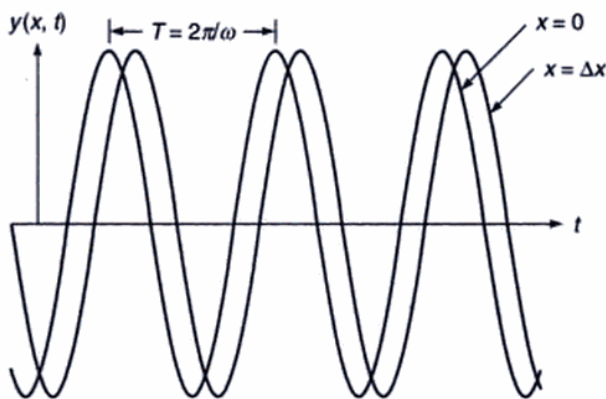


Fig. 9.5 The curves represent the time variation of the displacement at $x = 0$ at $x = \Delta x$ respectively when a sinusoidal wave is propagating the $+x$ -direction.

where

$$\omega = kv \quad (10)$$

The curves correspond to the time variation of the displacement of the two points. Corresponding to a particular point, the displacement repeats itself after a time

$$T = 2\pi/\omega \quad (11)$$

which is known as the time period of the wave. The quantity

$$\nu = 1/T \quad (12)$$

is known as the frequency of the wave and represents the number of oscillations that a particle carries out in one second. It can be seen from the two curves in Fig. 9.5, that the two points $x = 0$ and $x = \Delta x$ execute exactly similar vibrations except for a phase difference of $k\Delta x$. In fact any two points on the string execute simple harmonic motions with the same amplitude and same frequency but with a phase difference of kx_0 where x_0 represents the distance between the two points. Clearly if this distance is a multiple of the wavelength, i.e.,

$$x_0 = m\lambda, m = 1, 2, \dots$$

then

$$kx_0 = \frac{2\pi}{\lambda} m\lambda = 2m\pi$$

which implies that two points separated by a distance which is a multiple of the wavelength vibrate with the same phase.

Similarly, two points separated by a distance $\frac{1}{2}\lambda, \frac{3}{2}\lambda, \dots$ vibrate in opposite phase. In general, a path difference of x_0 corresponds to a phase difference of $\frac{2\pi}{\lambda} x_0$.

Using Eqs (10)–(12), we get

$$\nu = \frac{1}{T} = \frac{\omega}{2\pi} = \frac{kv}{2\pi} = \frac{v}{\lambda}$$

or

$$v = \nu\lambda \quad (13)$$

Notice the similarity in the variation of the displacement with respect to x (at a given value of time) and with respect to t (at a given value of x); see Figs 9.4 and 9.5. The similarity can be expressed by writing Eq. (6) in the form

$$y(x, t) = a \cos \left(\frac{2\pi}{\lambda} x - \frac{2\pi}{T} t \right) \quad (14)$$

which shows that the wavelength λ in Fig. 9.4 plays the same role as the time period T in Fig. 9.5. Equation (14) is often written in the form

$$y(x, t) = a \cos (kx - \omega t) \quad (15)$$

It should be pointed out that the entire discussion given above would remain valid for an arbitrary value of the phase factor ϕ .

9.3 TYPES OF WAVES

As mentioned earlier, when a wave is propagating through a string the displacement is at right angles to the direction of propagation. Such a wave is known as a transverse wave.* Similarly, when a sound wave propagates through air the displacement of the air molecules are along the direction of propagation of the wave; such waves are known as longitudinal waves. However, there are waves which are neither longitudinal nor transverse in character; for example, when a wave propagates through the surface of water, the water molecules move approximately in circular orbits.

9.4 ENERGY TRANSPORT IN WAVE MOTION

A wave carries energy; for example, when a transverse wave propagates through a string, the particles execute simple harmonic motions about their equilibrium positions and associated with this motion is a certain amount of energy. As the wave propagates through, the energy gets transported from one end of the string to the other. We consider the time variation of the displacement of a particle, which can be written as

$$y = a \cos (\omega t + \phi) \quad (16)$$

*Electromagnetic waves are also transverse in character. However, it should be mentioned that the electromagnetic waves have also a longitudinal component near the source which dies off rapidly at large distances (see Sec. 20.4).

The instantaneous velocity of the particle would be

$$v = \frac{dy}{dt} = -a\omega \sin(\omega t + \phi) \quad (17)$$

Thus, the kinetic energy (T) would be given by

$$\begin{aligned} T &= \frac{1}{2} m \left(\frac{dy}{dt} \right)^2 \\ &= \frac{1}{2} m a^2 \omega^2 \sin^2(\omega t + \phi) \end{aligned} \quad (18)$$

The total energy (E) will be the maximum value of T

$$\begin{aligned} E &= (T)_{\max} \\ &= \frac{1}{2} m a^2 \omega^2 [\sin^2(\omega t + \phi)]_{\max} \\ &= \frac{1}{2} m a^2 \omega^2 \end{aligned} \quad (19)$$

For a sound wave propagating through a gas, the energy per unit volume, ϵ , would be given by

$$\begin{aligned} \epsilon &= \frac{1}{2} m n a^2 \omega^2 \\ &= \frac{1}{2} \rho a^2 \omega^2 \\ &= 2\pi^2 \rho a^2 v^2 \end{aligned} \quad (20)$$

where n represents the number of molecules per unit volume and $\rho (= nm)$ the density of the gas. With such a wave, we can associate the intensity which is defined as the energy flow per unit time across a unit area perpendicular to the direction of propagation. Since the speed of propagation of the wave is v , the intensity (I) would be given by*

$$I = 2\pi^2 \rho v a^2 v^2 \quad (21)$$

Thus the intensity is proportional to the square of the amplitude and square of the frequency.

Let us consider a wave emanating from a point source in a uniform isotropic** medium. We will assume that there is no absorption and that the source is emitting W joules per second (W represents the power of the source). Consider a sphere of radius r whose centre is at the point source. Clearly, W joules per second will cross the spherical surface whose area is $4\pi r^2$. Thus, the intensity I would be given by

$$I = \frac{W}{4\pi r^2} \quad (22)$$

*This can be easily understood from the fact that if we have N particles per unit volume, each moving with the same velocity v , then the number of particles crossing an unit area (normal to v) per unit time would be Nv .

**Isotropic media are the ones in which physical properties (like velocity of propagation of a particular wave) are the same in all directions. In Chapter 19 we will consider anisotropic media.

***When waves emanate from a point source in an isotropic medium, all the points on the surface of a sphere (whose centre is at the point source) have the same amplitude and the same phase; in other words, the locus of points which have the same amplitude and the same phase is a sphere. Such waves are known as spherical waves. Far away from the source, over a small area, the spherical waves are essentially plane waves.

which is nothing but the inverse square law. Using Eqs (21) and (22) we obtain

$$\frac{W}{4\pi r^2} = 2\pi^2 \rho v a^2 v^2$$

or

$$a = \left[\frac{W}{8\pi^3 \rho v v^2} \right]^{1/2} \frac{1}{r} \quad (23)$$

showing that the amplitude falls off as $\frac{1}{r}$. Indeed, for a spherical wave*** emanating from a point source, the displacement is given by

$$f = \frac{a_0}{r} \sin(kr - \omega t)$$

where a_0 represents the amplitude of the wave at unit distance from the source.

Example 9.3 A source of sound is vibrating with a frequency of 256 vibrations per second in air and propagating energy uniformly in all directions at the rate of 5 Joules per second. Calculate the intensity and the amplitude of the wave at a distance of 25 m from the source. Assume that there is no absorption [speed of sound waves in air = 330 m/sec; density of air = 1.29 kg/m³].

Solution:

$$\begin{aligned} \text{Intensity } I &= \frac{5 \text{ J/s}}{4\pi \times (25)^2 \text{ m}^2} \\ &= 6.3 \times 10^{-4} \text{ J sec}^{-1} \text{ m}^{-2} \end{aligned}$$

$$\begin{aligned} \text{Thus } a &= \left[\frac{5}{8\pi^3 \times 1.29 \times 330 \times 256 \times 256} \right]^{1/2} \frac{1}{25} \\ &\approx 3 \times 10^{-7} \text{ m.} \end{aligned}$$

Example 9.4 Show that when a transverse wave propagates through a string, the energy transmitted per unit time is $\frac{1}{2} \rho \omega^2 a^2 v$ where ρ is the mass per unit length, a the amplitude of the wave and v the speed of propagation of the wave.

Solution: The energy associated per unit length of the string is $\frac{1}{2} \rho \omega^2 a^2$; since the speed of the wave is v , the result follows.

9.5 THE ONE-DIMENSIONAL WAVE EQUATION

In Sec. 9.1 we had shown that the displacement ψ of a one-dimensional wave is always of the form

$$\psi = f(x - vt) + g(x + vt) \quad (24)$$

where the first term on the RHS of the above equation represents a disturbance propagating in the $+x$ direction with speed v and similarly, the second term represents a disturbance propagating in the $-x$ direction with speed v . The question now arises as to how we can predict the existence of waves and what would be the velocity of propagation of these waves? The answer to this question is as follows: If we can derive an equation of the form

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2} \quad (25)$$

from physical considerations, then we can be sure that waves will result and ψ will represent the displacement associated with the wave. This follows from the fact that the general solution of Eq. (25) is of the form

$$\psi = f(x - vt) + g(x + vt) \quad (26)$$

where f and g are arbitrary functions of their argument. Although the general solution of Eq. (25) will be derived in Sec. 9.9, we will show here that the solution expressed by Eq. (26) indeed satisfies Eq. (25). We rewrite Eq. (26) in the form

$$\psi = f(\xi) + g(\eta) \quad (27)$$

where

$$\xi = x - vt \quad (28a)$$

and

$$\eta = x + vt \quad (28b)$$

$$\text{Thus } \frac{\partial \psi}{\partial x} = f'(\xi) \frac{\partial \xi}{\partial x} + g'(\eta) \frac{\partial \eta}{\partial x}$$

where primes denote differentiation with respect to the argument, i.e.

$$f'(\xi) = \frac{df}{d\xi}, \text{ etc.}$$

$$\text{Since } \frac{\partial \xi}{\partial x} = 1 = \frac{\partial \eta}{\partial x} \text{ [see Eq. (28)], we obtain}$$

$$\frac{\partial \psi}{\partial x} = f'(\xi) + g'(\eta)$$

Differentiating once more, we get

$$\frac{\partial^2 \psi}{\partial x^2} = f''(\xi) \frac{\partial \xi}{\partial x} + g''(\eta) \frac{\partial \eta}{\partial x}$$

$$= f''(\xi) + g''(\eta) \quad (29)$$

Similarly,

$$\begin{aligned} \frac{\partial \psi}{\partial t} &= f'(\xi) \frac{\partial \xi}{\partial t} + g'(\eta) \frac{\partial \eta}{\partial t} \\ &= v [-f'(\xi) + g'(\eta)] \end{aligned}$$

and

$$\frac{\partial^2 \psi}{\partial t^2} = v^2 [f''(\xi) + g''(\eta)] \quad (30)$$

$$\text{Thus } \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2} = f''(\xi) + g''(\eta) = \frac{\partial^2 \psi}{\partial x^2} \quad (31)$$

which proves that the solution expressed by Eq. (26) indeed satisfies Eq. (25). Consequently, if we ever obtain an equation of the form of Eq. (25) from physical considerations, we can predict the existence of waves, the speed of which would be v .

We must mention that the simplest particular solutions of the wave equation correspond to sinusoidal variation:

$$\psi = A \sin [k(x \pm vt) + \phi] \quad (32a)$$

or

$$\psi = A \cos [k(x \pm vt) + \phi] \quad (32b)$$

As shown in Sec. 9.2,

$$k = 2\pi/\lambda \quad \text{and} \quad kv = \omega = 2\pi\nu$$

where λ is the wavelength and ν the frequency of the wave. Instead of sinusoidal variation it is often more convenient to write the solution in the form

$$\psi = A \exp [i(kx \pm \omega t + \phi)] \quad (33)$$

where, as before, A and ϕ represent the amplitude and initial phase of the wave. In writing Eq. (33), it is implied that the actual displacement is just the real part of ψ which is

$$A \cos (kx \pm \omega t + \phi)$$

In the next three sections we will derive the wave equation for some simple cases.* In Sec. 9.9 we will discuss the general solution of the wave equation.

9.6 TRANSVERSE VIBRATIONS OF A STRETCHED STRING

Let us consider a stretched string having a tension T . In its equilibrium position the string is assumed to lie on the x -axis. If the string is pulled in the y -direction then forces will act on the string which will tend to bring it back to its equilibrium position. Let us consider a small length AB of the string and calculate the net force acting on it in the

*In Chapter 20 we will derive the wave equation from Maxwell's equations and thereby obtain an expression for the speed of electromagnetic waves.

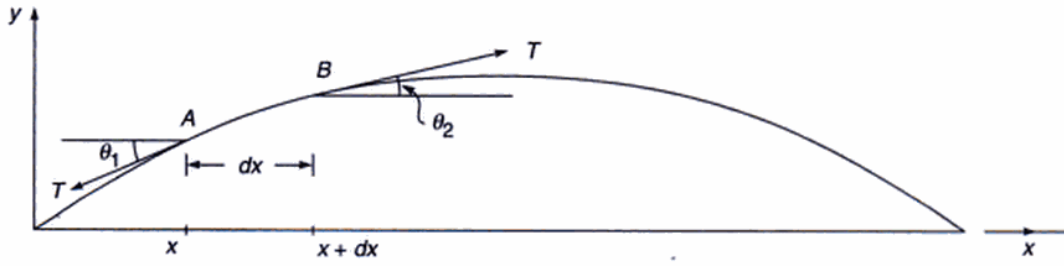


Fig. 9.6 Transverse vibrations of a stretched string.

y-direction. Due to the tension T , the end points A and B experience force in the direction of the arrows shown in Fig. 9.6. The force at A in the upward direction is

$$-T \sin \theta_1 \approx -T \tan \theta_1 = -T \left. \frac{\partial y}{\partial x} \right|_x$$

Similarly, the force at B in the upward direction is

$$T \sin \theta_2 \approx T \tan \theta_2 = T \left. \frac{\partial y}{\partial x} \right|_{x+dx}$$

where we have assumed θ_1 and θ_2 to be small. Thus the net force acting on AB in the y -direction is

$$T \left[\left(\frac{\partial y}{\partial x} \right)_{x+dx} - \left(\frac{\partial y}{\partial x} \right)_x \right] = T \frac{\partial^2 y}{\partial x^2} dx, \quad (34)$$

where we have used the Taylor series expansion of

$\left(\frac{\partial y}{\partial x} \right)_{x+dx}$ about the point x :

$$\left(\frac{\partial y}{\partial x} \right)_{x+dx} = \left(\frac{\partial y}{\partial x} \right)_x + \frac{\partial}{\partial x} \left(\frac{\partial y}{\partial x} \right) dx$$

and have neglected higher order terms because dx is infinitesimal. The equation of motion is therefore

$$\Delta m \frac{\partial^2 y}{\partial t^2} = T \frac{\partial^2 y}{\partial x^2} dx$$

where Δm is the mass of the element AB . If ρ is the mass per unit length, then

$$\Delta m = \rho dx$$

and we get

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{T/\rho} \frac{\partial^2 y}{\partial t^2} \quad (35)$$

which is the one-dimensional wave equation. Thus we may conclude that transverse waves can propagate through a stretched string and if we compare the above equation with Eq. (25) we obtain the following expression for the speed of the transverse waves:

$$v = \sqrt{T/\rho} \quad (36)$$

The vibrations of a clamped string will be discussed in Sec. 11.2. It should be mentioned that in an actual string, the displacement is not rigorously of the form given by Eq. (24); this is a consequence of the various approximations made in the derivation of the wave equation. There is, in general, an attenuation of the wave and also the shape does not remain unaltered.

9.7 LONGITUDINAL SOUND WAVES IN A SOLID

In this section we will derive an expression for the velocity of longitudinal sound waves propagating in an elastic solid. Let us consider a solid cylindrical rod of cross-sectional area A . Let PQ and RS be two transverse sections of the rod at distances x and $x + \Delta x$ from a fixed point O , where we have chosen the x -axis to be along the length of the rod (see Fig. 9.7).

Let the longitudinal displacement of a plane be denoted by $\xi(x)$. Thus the displacements of the planes PQ and RS would be $\xi(x)$ and $\xi(x + \Delta x)$ respectively. In the displaced position, the distance between the planes $P'Q'$ and $R'S'$ would be

$$\begin{aligned} \xi(x + \Delta x) - \xi(x) + \Delta x &= \xi(x) + \frac{\partial \xi}{\partial x} \Delta x - \xi(x) + \Delta x \\ &= \Delta x + \frac{\partial \xi}{\partial x} \Delta x \end{aligned}$$

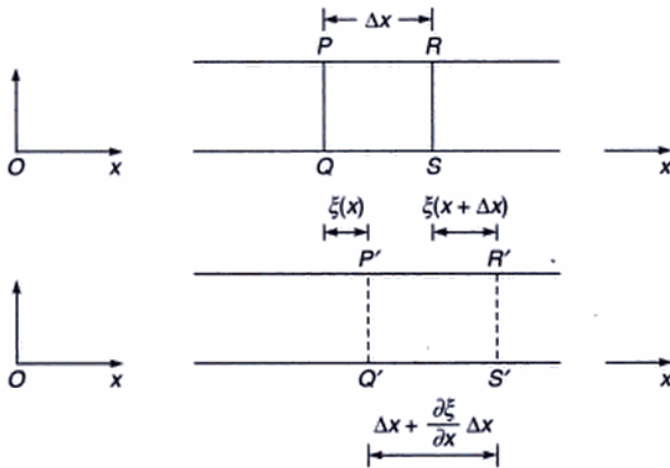


Fig. 9.7 Propagation of longitudinal sound waves through a cylindrical rod.

The elongation of the element would be $\frac{\partial \xi}{\partial x} \Delta x$ and, therefore, the longitudinal strain would be

$$\frac{\text{Increase in length}}{\text{Original length}} = \frac{\frac{\partial \xi}{\partial x} \Delta x}{\Delta x} = \frac{\partial \xi}{\partial x} \quad (37)$$

Since the Young's modulus (Y) is defined as the ratio of the longitudinal stress to the longitudinal strain, we have

$$\begin{aligned} \text{Longitudinal stress} &= \frac{F}{A} = Y \times \text{Strain} \\ &= Y \frac{\partial \xi}{\partial x} \end{aligned} \quad (38)$$

where F is the force acting on the element $P'Q'$. Thus

$$F(x) = YA \frac{\partial \xi}{\partial x} \quad (39)$$

and, therefore,

$$\frac{\partial F}{\partial x} = YA \frac{\partial^2 \xi}{\partial x^2} \quad (40)$$

Now, if we consider the volume $P'Q'S'R'$ then a force F is acting on the element $P'Q'$ in the negative x -direction and a force $F(x + \Delta x)$ is acting on the plane $R'S'$ along the positive x -direction. Thus the resultant force acting on the element $P'Q'S'R'$ will be

$$\begin{aligned} F(x + \Delta x) - F(x) &= \frac{\partial F}{\partial x} \Delta x \\ &= YA \frac{\partial^2 \xi}{\partial x^2} \Delta x \end{aligned} \quad (41)$$

If ρ represents the density, then the mass of the element would be $\rho A \Delta x$. Thus the equation of motion will be

$$\begin{aligned} \rho A \Delta x \frac{\partial^2 \xi}{\partial t^2} &= YA \Delta x \frac{\partial^2 \xi}{\partial x^2} \\ \text{or} \quad \frac{\partial^2 \xi}{\partial x^2} &= \frac{1}{v_l^2} \frac{\partial^2 \xi}{\partial t^2} \end{aligned} \quad (42)$$

$$\text{where} \quad v_l = \left(\frac{Y}{\rho} \right)^{1/2} \quad (43)$$

represents the velocity of the waves and the subscript l refers to the fact that we are considering longitudinal waves.*

The above derivation is valid when the transverse dimension of the rod is small compared with the wavelength of the disturbance so that one may assume that the longitudinal displacement at all points on any transverse section (like PQ) are the same. In general, if one carries out a rigorous analysis of the vibrations of an extended isotropic elastic solid, one can show that the velocities of the longitudinal and transverse waves will be given by**

$$v_l = \left(\frac{Y}{\rho} \frac{(1 - \sigma)}{(1 + \sigma)(1 - 2\sigma)} \right)^{1/2} = \left(\frac{K + \frac{4}{3}\eta}{\rho} \right)^{1/2} \quad (44)$$

$$v_t = \left(\frac{Y}{\rho} \frac{1}{2(1 + \sigma)} \right)^{1/2} = \left(\frac{\eta}{\rho} \right)^{1/2} \quad (45)$$

where σ , η and K represent the Poisson ratio, modulus of rigidity and bulk modulus respectively. We must mention that the transverse wave [whose velocity is given by Eq. (45)] is due to the restoring forces arising because of the elastic properties of the material, whereas corresponding to the transverse waves discussed in Sec. 9.6, the string moved as a whole and the restoring force was due to the externally applied tension.

*In a similar manner one can consider transverse waves propagating through an elastic solid, the velocity of which would be given by [see, for example, Ref. 8]

$$v_t = \sqrt{\eta/\rho}$$

where η represents the modulus of rigidity.

**See, for example, Ref. 5.

9.8 LONGITUDINAL WAVES IN A GAS

In order to determine the speed of propagation of longitudinal sound waves in a gas, we consider a column $PQSR$ as shown in Fig. 9.8 (a). Once again, because of a longitudinal displacement, the plane PQ gets displaced by $\xi(x)$ and the plane RS gets displaced by a distance $\xi(x + \Delta x)$ (see Fig. 9.8). Let the pressure of the gas in the absence of any disturbance be P_0 . Let $P_0 + \Delta P(x)$ and $P_0 + \Delta P(x + \Delta x)$ denote the pressures at the planes $P'Q'$ and $R'S'$ respectively. Now, if we consider the column $P'Q'S'R'$ then the pressure $P_0 + \Delta P(x)$ on the face $P'Q'$ acts in the $+x$ direction whereas the pressure $P_0 + \Delta P(x + \Delta x)$ on the face $R'S'$ acts in the $-x$ direction. Thus the force acting on the column $P'Q'S'R'$ would be

$$\begin{aligned} & [\Delta P(x) - \Delta P(x + \Delta x)]A \\ &= -\frac{\partial}{\partial x} (\Delta P) \Delta x A \end{aligned} \quad (46)$$

where A represents the cross-sectional area. Consequently, the equation of motion for the column $P'Q'S'R'$ would be

$$-\frac{\partial}{\partial x} (\Delta P) A \Delta x = \rho A \Delta x \frac{\partial^2 \xi}{\partial t^2}$$

where ρ represents the density of the gas. Thus

$$-\frac{\partial}{\partial x} (\Delta P) = \rho \frac{\partial^2 \xi}{\partial t^2} \quad (47)$$

Now, a change in pressure gives rise to a change in volume, and if the frequency of the wave is large (≥ 20 Hz), the

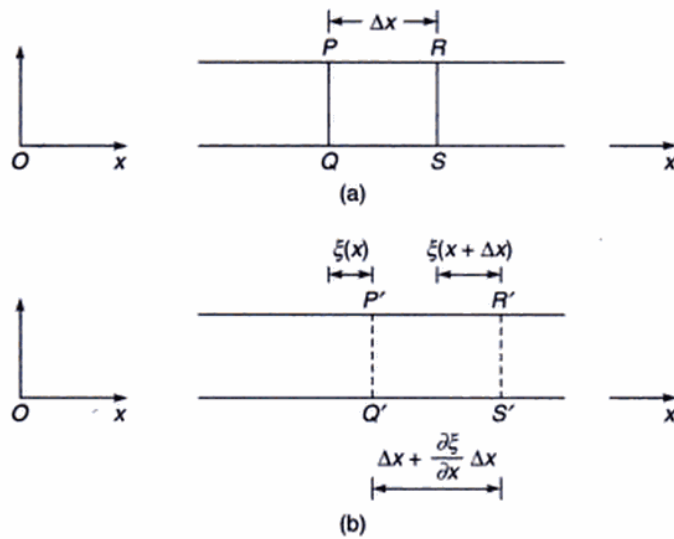


Fig. 9.8 Propagation of longitudinal sound waves through air.

pressure fluctuations will be rapid and one may assume the process to be adiabatic. Thus, we may write

$$PV^\gamma = \text{constant} \quad (48)$$

where $\gamma = C_p/C_v$ represents the ratio of the two specific heats. If we differentiate the above expression, we get

$$\begin{aligned} \Delta P V^\gamma + \gamma V^{\gamma-1} P \Delta V &= 0 \\ \Delta P &= -\frac{\gamma P}{V} \Delta V \end{aligned} \quad (49)$$

The change in the length of the column $PQSR$ is

$$[\xi(x + \Delta x) - \xi(x) + \Delta x] - \Delta x = \frac{\partial \xi}{\partial x} \Delta x$$

Thus, the change in the volume

$$\Delta V = \frac{\partial \xi}{\partial x} A \Delta x$$

The original volume V of the element is $A \Delta x$. Thus

$$\begin{aligned} \Delta P &= -\frac{\gamma P}{A \Delta x} \frac{\partial \xi}{\partial x} A \Delta x \\ &= -\gamma P \frac{\partial \xi}{\partial x} \end{aligned} \quad (50)$$

$$\text{or} \quad \frac{\partial}{\partial x} (\Delta P) = -\gamma P \frac{\partial^2 \xi}{\partial x^2} \quad (51)$$

Using Eqs. (47) and (51), we obtain

$$\frac{\partial^2 \xi}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 \xi}{\partial t^2} \quad (52)$$

where

$$v = \left(\frac{\gamma P}{\rho} \right)^{1/2} \quad (53)$$

represents the velocity of propagation of longitudinal sound waves in a gas. For air, if we assume $\gamma = 1.40$, $P = 1.01 \times 10^6$ dynes/cm² and $\rho = 1.3 \times 10^{-3}$ g/cm³, then we obtain

$$v \approx 330 \text{ m/sec}$$

The adiabatic compressibility of a gas is given by

$$\kappa_s = \frac{1}{V} \left(\frac{\partial V}{\partial P} \right)_s = \frac{1}{\gamma P} \quad (54)$$

where the subscript s refers to the adiabatic condition (constant entropy). The bulk modulus (K) of a gas is the inverse of κ_s .

$$K = \frac{1}{\kappa_s} = \gamma P \quad (55)$$

and if we substitute this expression for K in Eq. (44), we obtain Eq. (53) where we have used the fact that the modulus of rigidity (η) for a gas is zero.

9.9 THE GENERAL SOLUTION OF THE ONE-DIMENSIONAL WAVE EQUATION*

In order to obtain a general solution of the equation

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2} \quad (56)$$

we introduce two new variables

$$\xi = x - vt \quad (57)$$

$$\eta = x + vt \quad (58)$$

and write Eq. (56) in terms of these variables. Now,

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial \xi} \frac{\partial \xi}{\partial x} + \frac{\partial \psi}{\partial \eta} \frac{\partial \eta}{\partial x} \quad (59)$$

or

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial \xi} + \frac{\partial \psi}{\partial \eta} \quad (60)$$

where we have used the fact that

$$\frac{\partial \xi}{\partial x} = 1 \quad \text{and} \quad \frac{\partial \eta}{\partial x} = 1$$

Differentiating Eq. (60) with respect to x , we get

$$\begin{aligned} \frac{\partial^2 \psi}{\partial x^2} &= \frac{\partial}{\partial x} \left(\frac{\partial \psi}{\partial \xi} \right) + \frac{\partial}{\partial x} \left(\frac{\partial \psi}{\partial \eta} \right) \\ &= \frac{\partial}{\partial \xi} \left(\frac{\partial \psi}{\partial \xi} \right) \frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta} \left(\frac{\partial \psi}{\partial \xi} \right) \frac{\partial \eta}{\partial x} \\ &\quad + \frac{\partial}{\partial \xi} \left(\frac{\partial \psi}{\partial \eta} \right) \frac{\partial \xi}{\partial x} + \frac{\partial}{\partial \eta} \left(\frac{\partial \psi}{\partial \eta} \right) \frac{\partial \eta}{\partial x} \end{aligned}$$

or

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{\partial^2 \psi}{\partial \xi^2} + 2 \frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2} \quad (61)$$

Similarly

$$\frac{\partial \psi}{\partial t} = \frac{\partial \psi}{\partial \xi} \frac{\partial \xi}{\partial t} + \frac{\partial \psi}{\partial \eta} \frac{\partial \eta}{\partial t}$$

$$= -v \frac{\partial \psi}{\partial \xi} + v \frac{\partial \psi}{\partial \eta}$$

and

$$\begin{aligned} \frac{\partial^2 \psi}{\partial t^2} &= -v \left[\frac{\partial}{\partial \xi} \left(\frac{\partial \psi}{\partial \xi} \right) \frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta} \left(\frac{\partial \psi}{\partial \xi} \right) \frac{\partial \eta}{\partial t} \right] \\ &\quad + v \left[\frac{\partial}{\partial \xi} \left(\frac{\partial \psi}{\partial \eta} \right) \frac{\partial \xi}{\partial t} + \frac{\partial}{\partial \eta} \left(\frac{\partial \psi}{\partial \eta} \right) \frac{\partial \eta}{\partial t} \right] \end{aligned}$$

or

$$\frac{\partial^2 \psi}{\partial t^2} = v^2 \left[\frac{\partial^2 \psi}{\partial \xi^2} - 2 \frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2} \right] \quad (62)$$

Substituting the expressions for $\partial^2 \psi / \partial x^2$ and $\partial^2 \psi / \partial t^2$ from Eqs. (61) and (62) in Eq. (56) we obtain

$$\frac{\partial^2 \psi}{\partial \xi^2} + 2 \frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2} = \frac{\partial^2 \psi}{\partial \xi^2} - 2 \frac{\partial^2 \psi}{\partial \eta \partial \xi} + \frac{\partial^2 \psi}{\partial \eta^2}$$

or

$$\frac{\partial}{\partial \eta} \left(\frac{\partial \psi}{\partial \xi} \right) = 0 \quad (63)$$

Thus $\partial \psi / \partial \xi$ has to be independent of η ; however, it can be an arbitrary function of ξ :

$$\frac{\partial \psi}{\partial \xi} = F(\xi) \quad (64)$$

or

$$\psi = \int F(\xi) d\xi + \text{constant of integration.}$$

The constant of integration can be an arbitrary function of η and since the integral of an arbitrary function is again an arbitrary function, we obtain the following general solution of the wave equation

$$\begin{aligned} \psi &= f(\xi) + g(\eta) \\ &= f(x - vt) + g(x + vt) \end{aligned} \quad (65)$$

where f and g are arbitrary functions of their argument. The function $f(x - vt)$ represents a disturbance propagating in the $+x$ direction with speed v and the function $g(x + vt)$ represents a disturbance propagating in the $-x$ direction.

Example 9.5 Solve the one-dimensional wave equation [Eq. (25)] by the method of separation of variables** and show that the solution can indeed be expressed in the form given by Eqs. (32) and (33).

*This section may be skipped in the first reading.

**The method of separation of variables is a powerful method for solving certain kinds of partial differential equations. According to this method the solution is assumed to be a product of functions, each function depending only on one independent variable [see Eq. (67)]. On substituting this solution, if the variables separate out then the method is said to work and the general solution is a linear sum of all possible solutions; see, e.g., the analysis given in Sec. 7.2. If the variables do not separate out one has to try some other method to solve the equation.

Solution: In the method of separation of variables, we try a solution of the wave equation

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2} \quad (66)$$

of the form

$$\psi(x, t) = X(x) T(t) \quad (67)$$

where $X(x)$ is a function of x alone and $T(t)$ is a function of t alone. Substituting in Eq. (66), we get

$$T(t) \frac{d^2 X}{dx^2} = \frac{1}{v^2} X(x) \frac{d^2 T}{dt^2}$$

or*

$$\frac{1}{X(x)} \frac{d^2 X}{dx^2} = \frac{1}{v^2 T(t)} \frac{d^2 T}{dt^2} \quad (68)$$

The LHS is a function of x alone and the RHS is a function of t alone. This implies that a function of one independent variable x is equal to a function of another independent variable t for all values of x and t . This is possible only when each side is equal to a constant; we set this constant equal to $-k^2$, thus

$$\frac{1}{X(x)} \frac{d^2 X}{dx^2} = \frac{1}{v^2} \frac{1}{T(t)} \frac{d^2 T}{dt^2} = -k^2 \quad (69)$$

or

$$\frac{d^2 X}{dx^2} + k^2 X(x) = 0 \quad (70)$$

and

$$\frac{d^2 T}{dt^2} + \omega^2 T(t) = 0 \quad (71)$$

where

$$\omega = kv = \frac{2\pi v}{\lambda} \quad (72)$$

represents the angular frequency of the wave. The solutions of Eqs (70) and (71) can easily be written down:

$$X(x) = (A \cos kx + B \sin kx)$$

and

$$T(t) = (C \cos \omega t + D \sin \omega t)$$

Thus

$$\psi(x, t) = (A \cos kx + B \sin kx) (C \cos \omega t + D \sin \omega t) \quad (73)$$

Suitable choice of the constants A , B , C and D would give

$$\psi(x, t) = a \cos (kx - \omega t + \phi)$$

or

$$\psi(x, t) = a \cos (kx + \omega t + \phi)$$

representing waves propagating in the $+x$ and $-x$ directions respectively. One can also have

$$\psi(x, t) = a \exp [\pm i(kx \pm \omega t + \phi)]$$

as a solution.

In general, all values of the frequencies are possible, but the frequency and wavelength have to be related through Eq. (72). However, there are systems (like a string under tension and fixed at both ends) where only certain values of frequencies are possible (see Sec. 7.2).

Example 9.6 Till now we have confined our discussion to waves in one dimension. The three-dimensional wave equation is of the form

$$\nabla^2 \psi = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2} \quad (74)$$

where

$$\nabla^2 \psi \equiv \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} + \frac{\partial^2 \psi}{\partial z^2} \quad (75)$$

Solve the three-dimensional wave equation by the method of separation of variables and interpret the solution physically.

Solution: Using the method of separation of variables, we write

$$\psi(x, y, z, t) = X(x) Y(y) Z(z) T(t) \quad (76)$$

where $X(x)$ is a function of x alone, etc. Substituting in Eq. (74), we obtain

$$YZT \frac{d^2 X}{dx^2} + XZT \frac{d^2 Y}{dy^2} + XYT \frac{d^2 Z}{dz^2} = \frac{1}{v^2} XYZ \frac{d^2 T}{dt^2}$$

or dividing throughout by ψ

$$\left[\frac{1}{X} \frac{d^2 X}{dx^2} \right] + \left[\frac{1}{Y} \frac{d^2 Y}{dy^2} \right] + \left[\frac{1}{Z} \frac{d^2 Z}{dz^2} \right] = \frac{1}{v^2} \left[\frac{1}{T} \frac{d^2 T}{dt^2} \right] \quad (77)$$

Since the first term on the LHS is a function of x alone, the second term is a function of y alone, etc., each term must be set equal to a constant. We write

$$\begin{aligned} \frac{1}{X} \frac{d^2 X}{dx^2} &= -k_x^2 \\ \frac{1}{Y} \frac{d^2 Y}{dy^2} &= -k_y^2 \\ \frac{1}{Z} \frac{d^2 Z}{dz^2} &= -k_z^2 \end{aligned} \quad (78)$$

*Notice that partial derivatives have been replaced by total derivatives.

where k_x^2 , k_y^2 and k_z^2 are constants. Thus

$$\frac{1}{v^2} \left[\frac{1}{T} \frac{d^2 T}{dt^2} \right] = -(k_x^2 + k_y^2 + k_z^2)$$

or

$$\frac{d^2 T}{dt^2} + \omega^2 T(t) = 0 \quad (79)$$

where

$$\omega^2 = k^2 v^2 \quad (80)$$

and

$$k^2 = k_x^2 + k_y^2 + k_z^2$$

The solutions of Eqs. (78) and (79) could be written in terms of sine and cosine functions; it is more convenient to write them in terms of the exponentials:

$$\begin{aligned} \psi &= A \exp [i(k_x x + k_y y + k_z z \pm \omega t + \phi)] \\ &= A \exp [i(\mathbf{k} \cdot \mathbf{r} \pm \omega t + \phi)] \end{aligned} \quad (81)$$

where the vector \mathbf{k} is defined such that its x , y and z -components are k_x , k_y and k_z respectively. One could have also written

$$\psi = A \cos (\mathbf{k} \cdot \mathbf{r} - \omega t + \phi) \quad (82)$$

Consider a vector \mathbf{r} which is normal to \mathbf{k} ; thus $\mathbf{k} \cdot \mathbf{r} = 0$; consequently at a given time the phase of the disturbance is constant on a plane normal to \mathbf{k} . The direction of propagation of the disturbance is along \mathbf{k} and the phase fronts are planes normal to \mathbf{k} ; such waves are known as plane waves (see Fig. 9.9). Notice that for a given value of the frequency, the value of k^2 is fixed [see Eq. (80)]; however, we can have waves propagating in different directions depending on the values of k_x , k_y and k_z . For example, if

$$k_x = k \quad \text{and} \quad k_y = k_z = 0 \quad (83a)$$

we have a wave propagating along the x -axis, the phase fronts are parallel to the y - z plane. Similarly, for

$$k_x = \frac{k}{\sqrt{2}}, k_y = \frac{k}{\sqrt{2}}, k_z = 0 \quad (83b)$$

the waves are propagating in a direction which makes equal angles with x - and y -axes (see Fig. 9.9).

Example 9.7 For a spherical wave the displacement ψ depends only on r and t where r is the magnitude of the

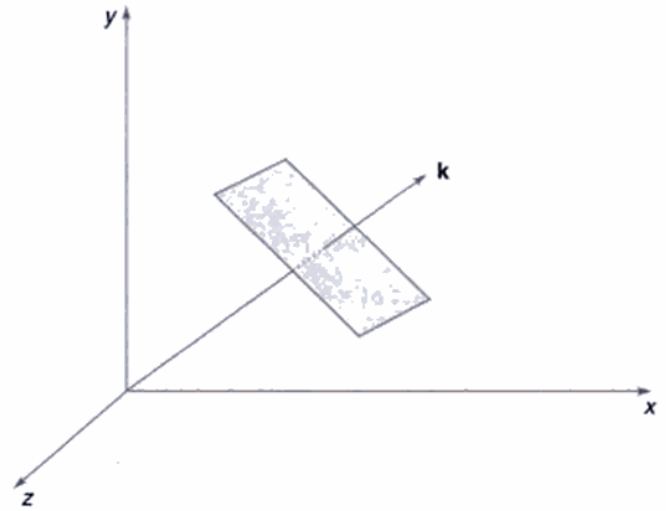


Fig. 9.9 Propagation of a plane wave along the direction \mathbf{k} . $\left(k_x = k_y = \frac{k}{\sqrt{2}}, k_z = 0 \right)$

distance from a fixed point. Obtain a general solution of the wave equation for a spherical wave.

Solution: We will first show that for a spherical wave

$$\nabla^2 \psi = \frac{\partial^2 \psi}{\partial r^2} + \frac{2}{r} \frac{\partial \psi}{\partial r} = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \psi}{\partial r} \right) \quad (84)$$

Now*

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial r} \frac{\partial r}{\partial x} \quad (85)$$

Since

$$r^2 = x^2 + y^2 + z^2 \quad (86)$$

therefore,

$$2r \frac{\partial r}{\partial x} = 2x$$

or

$$\frac{\partial r}{\partial x} = \frac{x}{r}$$

Thus

$$\frac{\partial \psi}{\partial x} = \frac{x}{r} \frac{\partial \psi}{\partial r}$$

and

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{r} \frac{\partial \psi}{\partial r} + x \frac{\partial}{\partial r} \left(\frac{1}{r} \frac{\partial \psi}{\partial r} \right) \frac{\partial r}{\partial x}$$

*In general ψ is a function of the three spherical polar coordinates r , θ and ϕ and

$$\frac{\partial \psi}{\partial x} = \frac{\partial \psi}{\partial r} \frac{\partial r}{\partial x} + \frac{\partial \psi}{\partial \theta} \frac{\partial \theta}{\partial x} + \frac{\partial \psi}{\partial \phi} \frac{\partial \phi}{\partial x}$$

However, for a spherical wave, ψ is a function of r alone and the last two terms vanish.

$$= \frac{1}{r} \frac{\partial \psi}{\partial r} + \frac{x^2}{r} \left[\frac{1}{r} \frac{\partial^2 \psi}{\partial r^2} - \frac{1}{r^2} \frac{\partial \psi}{\partial r} \right]$$

or

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{r} \frac{\partial \psi}{\partial r} + \frac{x^2}{r^2} \frac{\partial^2 \psi}{\partial r^2} - \frac{x^2}{r^3} \frac{\partial \psi}{\partial r}$$

Similarly

$$\frac{\partial^2 \psi}{\partial y^2} = \frac{1}{r} \frac{\partial \psi}{\partial r} + \frac{y^2}{r^2} \frac{\partial^2 \psi}{\partial r^2} - \frac{y^2}{r^3} \frac{\partial \psi}{\partial r}$$

and

$$\frac{\partial^2 \psi}{\partial z^2} = \frac{1}{r} \frac{\partial \psi}{\partial r} + \frac{z^2}{r^2} \frac{\partial^2 \psi}{\partial r^2} - \frac{z^2}{r^3} \frac{\partial \psi}{\partial r}$$

Adding, we get

$$\nabla^2 \psi = \frac{3}{r} \frac{\partial \psi}{\partial r} + \frac{\partial^2 \psi}{\partial r^2} - \frac{1}{r} \frac{\partial \psi}{\partial r}$$

where we have used Eq. (86). Thus

$$\nabla^2 \psi = \frac{\partial^2 \psi}{\partial r^2} + \frac{2}{r} \frac{\partial \psi}{\partial r} = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \psi}{\partial r} \right) \quad (87)$$

Thus, the wave equation for a spherical wave simplifies to

$$\nabla^2 \psi = \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \psi}{\partial r} \right) = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2} \quad (88)$$

If we make the substitution

$$\psi = \frac{u(r, t)}{r}$$

then

$$\begin{aligned} \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial \psi}{\partial r} \right) &= \frac{1}{r^2} \frac{\partial}{\partial r} \left(r \frac{\partial u}{\partial r} - u \right) \\ &= \frac{1}{r} \frac{\partial^2 u}{\partial r^2} \end{aligned}$$

Thus Eq. (88) becomes

$$\frac{1}{r} \frac{\partial^2 u}{\partial r^2} = \frac{1}{v^2} \frac{1}{r} \frac{\partial^2 u}{\partial t^2}$$

or

$$\frac{\partial^2 u}{\partial r^2} = \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2} \quad (89)$$

which is of the same form as the one-dimensional wave equation. The general solution of Eq. (89) is therefore given by

$$\psi = \frac{f(r - vt)}{r} + \frac{g(r + vt)}{r} \quad (90)$$

the first and the second terms (on the RHS) representing an outgoing spherical wave and an incoming spherical wave respectively. For time dependence of the form $\exp(\pm i\omega t)$ one obtains

$$\psi = \frac{A}{r} \exp[i(kr \pm \omega t)] \quad (91)$$

Notice that the factor $1/r$ term implies that the amplitude of a spherical wave decreases inversely with r , and, therefore, the intensity will fall off as $1/r^2$.

Example 9.8 As mentioned earlier, the solution obtained by the method of separation of variables represents the general solution. Show that a pulse of the form

$$\psi(x, t) = A \exp \left(-\frac{(x - vt)^2}{\sigma^2} \right) \quad (92)$$

can be expressed as a superposition of the solutions obtained by the method of separation of variables.

Solution: An arbitrary disturbance propagating in the $+x$ direction can be written as a superposition of the functions

$$\cos k(x - vt) \quad \text{and} \quad \sin k(x - vt)$$

Thus

$$\begin{aligned} \psi(x, t) &= \int_0^\infty a(k) \cos k(x - vt) dk \\ &\quad + \int_0^\infty b(k) \sin k(x - vt) dk \end{aligned} \quad (93)$$

where the functions $a(k)$ and $b(k)$ are to be determined from the form of $\psi(x, t = 0)$. Now,

$$\begin{aligned} \psi(x, 0) &= \int_0^\infty a(k) \cos kx dk \\ &\quad + \int_0^\infty b(k) \sin kx dk \end{aligned} \quad (94)$$

If $\psi(x, 0)$ is an even function of x , $b(k) = 0$ and if $\psi(x, 0)$ is an odd function of x , $a(k) = 0$. (If $\psi(x, 0)$ is neither even nor odd, both $a(k)$ and $b(k)$ will be finite). In the present case

$$\psi(x, 0) = A \exp \left(-\frac{x^2}{\sigma^2} \right) \quad (95)$$

which is an even function of x . Thus

$$\psi(x, 0) = A \exp \left(-\frac{x^2}{\sigma^2} \right) = \int_0^\infty a(k) \cos kx dk \quad (96)$$

Using the Fourier cosine transform,

$$\begin{aligned} a(k) &= \frac{2}{\pi} \int_0^{\infty} \psi(x, 0) \cos kx \, dx \\ &= \frac{2}{\pi} A \int_0^{\infty} \exp\left(-\frac{x^2}{\sigma^2}\right) \cos kx \, dx \\ &= \frac{\sigma}{\sqrt{\pi}} A \exp\left(-\frac{k^2 \sigma^2}{4}\right) \end{aligned}$$

Hence

$$\begin{aligned} A \exp\left[-\frac{(x-vt)^2}{\sigma^2}\right] \\ = \frac{\sigma A}{\sqrt{\pi}} \int_0^{\infty} \exp\left(-\frac{k^2 \sigma^2}{4}\right) \cos(kx - \omega t) \, dk \quad (97) \end{aligned}$$

SUMMARY

- For a sinusoidal wave, the displacement is given by

$$\Psi = a \cos [kx \pm \omega t + \phi]$$

where a represents the amplitude of the wave, $\omega (= 2\pi\nu)$ the angular frequency of the wave, $k (= 2\pi/\lambda)$ the wave number and λ represents the wavelength associated with the wave. The upper and lower signs correspond to waves propagating in the $-x$ and $+x$ directions respectively. Such a displacement is indeed produced in a long stretched string at the end of which a continuously vibrating tuning fork is placed. The quantity ϕ is known as the phase of the wave.

- The most general solution of the wave equation

$$\frac{\partial^2 \psi}{\partial x^2} = \frac{1}{v^2} \frac{\partial^2 \psi}{\partial t^2}$$

is of the form

$$\psi = f(x - vt) + g(x + vt)$$

where f and g are arbitrary functions of their argument. The first term on the RHS of the above equation represents a disturbance propagating in the $+x$ direction with speed v and similarly, the second term represents a disturbance propagating in the $-x$ direction with speed v . Thus if we can derive the wave equation from physical considerations, then we can be sure that waves will result and ψ will represent the displacement associated with the wave.

- For a spherical wave, the displacement is given by

$$\Psi = \frac{A}{r} e^{i(kr \pm \omega t)}$$

where the $+$ and $-$ signs correspond to incoming and outgoing waves respectively. Notice that the factor $1/r$ term implies that the amplitude of a spherical wave decreases inversely with r , and therefore, the intensity will fall off as $1/r^2$.

PROBLEMS

- 9.1 The displacement associated with a wave is given by

- $y(x, t) = 0.1 \cos(0.2x - 2t)$
- $y(x, t) = 0.2 \sin(0.5x + 3t)$
- $y(x, t) = 0.5 \sin 2\pi(0.1x - t)$

where in each case x and y are measured in centimetres and t in seconds. Calculate the wavelength, amplitude, frequency and the velocity in each case.

- 9.2 A transverse wave ($\lambda = 15$ cm, $\nu = 200$ sec⁻¹) is propagating on a stretched string in the $+x$ -direction with an amplitude of 0.5 cm. At $t = 0$, the point $x = 0$ is at its equilibrium position moving in the upward direction. Write the equation describing the wave and if $\rho = 0.1$ g/cm, calculate the energy associated with the wave per unit length of the wire.

- 9.3 Assuming that the human ear can hear in the frequency range $20 < \nu < 20,000$ Hz, what will be the corresponding wavelength range?

- 9.4 Calculate the speed of longitudinal waves at NTP in (a) argon ($\gamma = 1.67$), (b) Hydrogen ($\gamma = 1.41$).

[Ans: 308 m/sec, 169 m/sec]

- 9.5 Consider a wave propagating in the $+x$ -direction with speed 100 cm/sec. The displacement at $x = 10$ cm is given by the following equation:

$$y(x = 10, t) = 0.5 \sin(0.4t)$$

where x and y are measured in centimetres and t in seconds. Calculate the wavelength and the frequency associated with the wave and obtain an expression for the time variation of the displacement at $x = 0$.

- 9.6 Consider a wave propagating in the $-x$ -direction whose frequency is 100 sec⁻¹. At $t = 5$ sec the displacement associated with the wave is given by the following equation:

$$y(x, t = 5) = 0.5 \cos(0.1x)$$

where x and y are measured in centimetres and t in seconds. Obtain the displacement (as a function of x) at $t = 10$ sec. What is the wavelength and the velocity associated with the wave?

- 9.7 Repeat Problem 9.6 corresponding to

$$y(x, t = 5) = 0.5 \cos(0.1x) + 0.4 \sin(0.1x + \pi/3)$$

- 9.8 A Gaussian pulse is propagating in the $+x$ -direction and at $t = t_0$ the displacement is given by

$$y(x, t = t_0) = a \exp \left[-\frac{(x - b)^2}{\sigma^2} \right]$$

Find $y(x, t)$.

- 9.9 A sonometer wire is stretched with a tension of 1 N. Calculate the velocity of transverse waves if $\rho = 0.2$ g/cm.

- 9.10 The displacement associated with a three-dimensional wave is given by

$$\psi(x, y, z, t) = a \cos \left[\frac{\sqrt{3}}{2} kx + \frac{1}{2} ky - \omega t \right]$$

Show that the wave propagates along a direction making an angle 30° with the x -axis.

- 9.11 Obtain the unit vector along the direction of propagation for a wave, the displacement of which is given by

$$\psi(x, y, z, t) = a \cos [2x + 3y + 4z - 5t]$$

where x, y and z are measured in centimetres and t in seconds. What will be the wavelength and the frequency of the wave?

$$\left[\text{Ans: } \frac{2}{\sqrt{29}} \hat{x} + \frac{3}{\sqrt{29}} \hat{y} + \frac{4}{\sqrt{29}} \hat{z} \right]$$

- 9.12 Calculate the velocity of longitudinal elastic waves in aluminium ($Y = 0.70 \times 10^{12}$ dynes/cm², $\rho = 2.7$ g/cm³). Compare this value of velocity with the speed of sound in air.

REFERENCES AND SUGGESTED READINGS

1. H. J. J. Braddick, *Vibrations, Waves and Diffraction*, McGraw-Hill Publishing Co., London, 1965.
2. F. S. Crawford, *Waves and Oscillations, Berkeley Physics Course*, Vol. III, McGraw-Hill Book Co., New York, 1968.
3. C. A. Coulson, *Waves*, Seventh Edition, Oliver & Boyd Ltd., Edinburgh, 1955.
4. W. C. Elmore and M. A. Heald, *Physics of Waves*, McGraw-Hill Publishing Co., Maidenhead, 1969.
5. G. Joos, *Theoretical Physics* (translated by I. M. Freeman), Blackie & Son Ltd., London, 1955.
6. H. J. Pain, *The Physics of Vibrations and Waves*, John Wiley & Sons, London, 1968.
7. Physical Science Study Committee, *Physics*, D. C. Heath and Co., Boston, Mass., 1967.
8. J. C. Slater and N. H. Frank, *Electromagnetism*, Dover Publications, New York, 1969.
9. R. A. Waldson, *Waves and Oscillations*, Van Nostrand Publishing Co., New York, 1964.

Chapter 10

Huygens' Principle and Its Applications

*Christiaan Huygens, a Dutch physicist, in a communication to the Academie des Science in Paris, propounded his wave theory of light (published in his *Traite de Lumiere* in 1690). He considered that light is transmitted through an all-pervading aether that is made up of small elastic particles, each of which can act as a secondary source of wavelets. On this basis, Huygens explained many of the known propagation characteristics of light, including the double refraction in calcite discovered by Bartholinus.*

—From the Internet

10.1 INTRODUCTION

The wave theory of light was first put forward by Christiaan Huygens in 1678. During that period, everyone believed in Newton's corpuscular theory, which had satisfactorily explained the phenomena of reflection, refraction, the rectilinear propagation of light and the fact that light could propagate through vacuum. So empowering was Newton's authority that the scientists around Newton believed in the corpuscular theory much more than Newton himself; as such, when Huygens put forward his wave theory, no one really believed him. On the basis of his wave theory, Huygens explained satisfactorily the phenomena of reflection, refraction and total internal reflection and also provided a simple explanation of the then recently discovered birefringence (see Ch. 19). As we will see later, Huygens' theory predicted that the velocity of light in a medium (like water) shall be less than the velocity of light in free space, which is just the converse of the prediction made from Newton's corpuscular theory (see Sec. 1.2).

The wave character of light was not really accepted until the interference experiments of Young and Fresnel (in the early part of the nineteenth century) which could only be explained on the basis of a wave theory. At a later date, the data on the speed of light through transparent media were also available which was consistent with the results obtained by using the wave theory. It should be pointed out that Huygens did not know whether the light waves were longitudinal or transverse and also how they propagate through vacuum. It was only in the later part of the nineteenth century, when Maxwell propounded his famous electromagnetic theory, could the nature of light waves be understood properly.

10.2 HUYGENS' THEORY

Huygens' theory is essentially based on a geometrical construction which allows us to determine the shape of the wavefront at any time, if the shape of the wavefront at an earlier time is known. A wavefront is the locus of the points which are in the same phase; for example, if we drop a small stone in a calm pool of water, circular ripples spread out from the point of impact, each point on the circumference of the circle (whose center is at the point of impact) oscillates with the same amplitude and same phase and thus we have a circular wavefront. On the other hand, if we have a point source emanating waves in a uniform isotropic medium, the locus of points which have the same amplitude and are in the same phase are spheres. In this case we have spherical wavefronts as shown in Fig. 10.1(a). At large distances from the source, a small portion of the sphere can be considered as a plane and we have what is known as a plane wave [see Fig. 10.1(b)].

Now, according to Huygens' principle, each point of a wavefront is a source of secondary disturbance and the wavelets emanating from these points spread out in all directions with the speed of the wave. The envelope of these wavelets gives the shape of the new wavefront. In Fig. 10.2, S_1S_2 represents the shape of the wavefront (emanating from the point O) at a particular time which we denote as $t = 0$. The medium is assumed to be homogeneous and isotropic, i.e., the medium is characterized by the same property at all points and the speed of propagation of the wave is the same in all directions. Let us suppose we want to determine the shape of the wavefront after a time interval of Δt . Then, with each point on the wavefront as center, we draw spheres of radius $v \Delta t$, where v is the speed of the wave in that

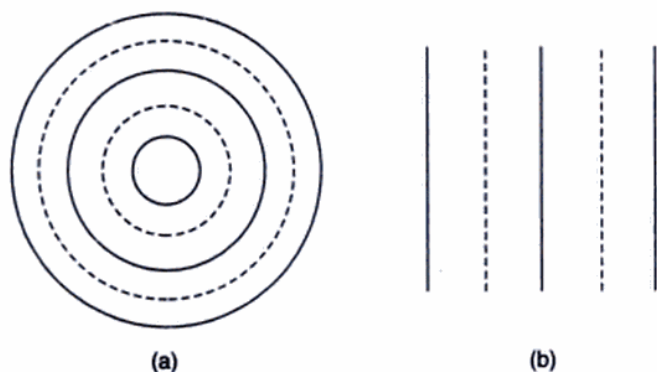


Fig. 10.1 (a) A point source emitting spherical waves. (b) At large distances, a small portion of the spherical wavefront can be approximated to a plane wavefront thus resulting in plane waves.

medium. If we draw a common tangent to all these spheres, then we obtain the envelope which is again a sphere centered at O . Thus the shape of the wavefront at a later time Δt is the sphere $S'_1S'_2$.

There is, however, one drawback with the above model, because we also obtain a backwave which is not present in practice. This backwave is shown as $S''_1S''_2$ in Fig. 10.2. In Huygens' theory, the presence of the backwave is avoided by assuming that the amplitude of the secondary wavelets is

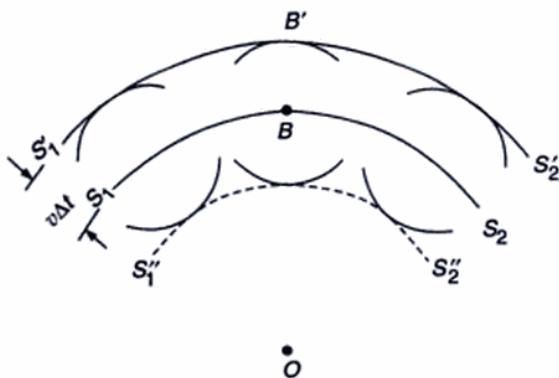


Fig. 10.2 Huygens' construction for the determination of the shape of the wavefront, given the shape of the wavefront at an earlier time. S_1S_2 is a spherical wavefront centered at O at a time, say $t = 0$. $S'_1S'_2$ corresponds to the state of the wavefront at a time Δt , which is again spherical and centered at O . The dashed curve represents the backwave.

not uniform in all directions; it is maximum in the forward direction and zero in the backward direction*. The absence of the backwave is really justified through the more rigorous wave theory.

In the next section we will discuss the original argument of Huygens to explain the rectilinear propagation of light. In Sec. 10.4 we will derive the laws of refraction and reflection by using Huygens' principle. Finally, in Sec. 10.5 we will show how Huygens' principle can be used in inhomogeneous media.

10.3 RECTILINEAR PROPAGATION

Let us consider spherical waves emanating from the point source O and striking the obstacle A (see Fig. 10.3). According to the rectilinear propagation of light (which is also predicted by corpuscular theory) one should obtain a shadow in the region PQ of the screen. As we will see in a later chapter, this is not rigorously true and one does obtain a finite intensity in the region of the geometrical shadow. However, at the time of Huygens, light was known to travel in straight lines and Huygens explained this by assuming that the secondary wavelets do not have any amplitude at any point not enveloped by the wavefront. Thus, referring back to Fig. 10.2, the secondary wavelets emanating from a

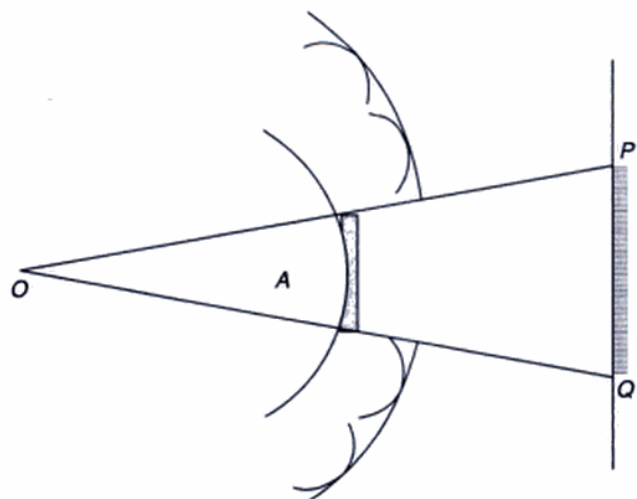


Fig. 10.3 Rectilinear propagation of light. O is a point source emitting spherical waves and A is an obstacle which forms a shadow in the region PQ of the screen.

* Indeed it can be shown from diffraction theory that one does obtain (under certain approximations) an obliquity factor, which is of the form $\frac{1}{2}(1 + \cos \theta)$ where θ is the angle between the normal to the wavefront and the direction under consideration. Clearly when $\theta = 0$, the obliquity factor is 1 (thereby giving rise to maximum amplitude in the forward direction) and when $\theta = \pi$, the obliquity factor is zero (thereby giving rise to zero amplitude in the backward direction).

typical point B will give rise to a finite amplitude at B' only and not at any other point.

The above explanation of the rectilinear propagation of light is indeed unsatisfactory and is incorrect. Further, as pointed out earlier, one does observe a finite intensity of light in the geometrical shadow. A satisfactory explanation was put forward by Fresnel, who postulated that the secondary wavelets mutually interfere. The Huygens' principle along with the fact that the secondary wavelets mutually interfere, is known as the Huygens-Fresnel principle. It may be mentioned that if a plane wave is allowed to fall on a tiny hole,* then the hole approximately acts as a point source and spherical waves emanate from it (see Figs. 10.4(a) and (b)). This fact is in direct contradiction to the original proposition of Huygens** according to which the secondary wavelets do not have any amplitude at any point not enveloped by the wavefront; however, as we will see in

the chapter on diffraction, it can be explained satisfactorily on the basis of Huygens-Fresnel principle.

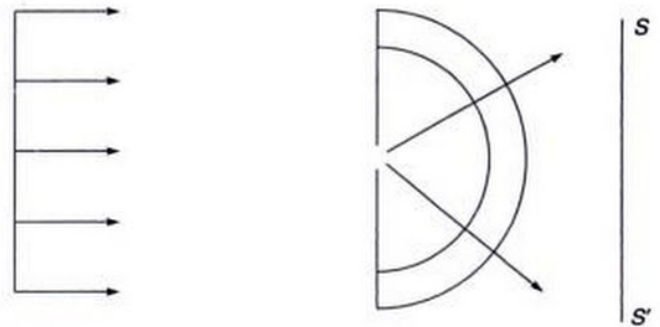


Fig. 10.4 (a) A plane wavefront is incident on a pin hole. If the diameter of the pinhole is small (compared to the wavelength) the entire screen SS' will be illuminated.

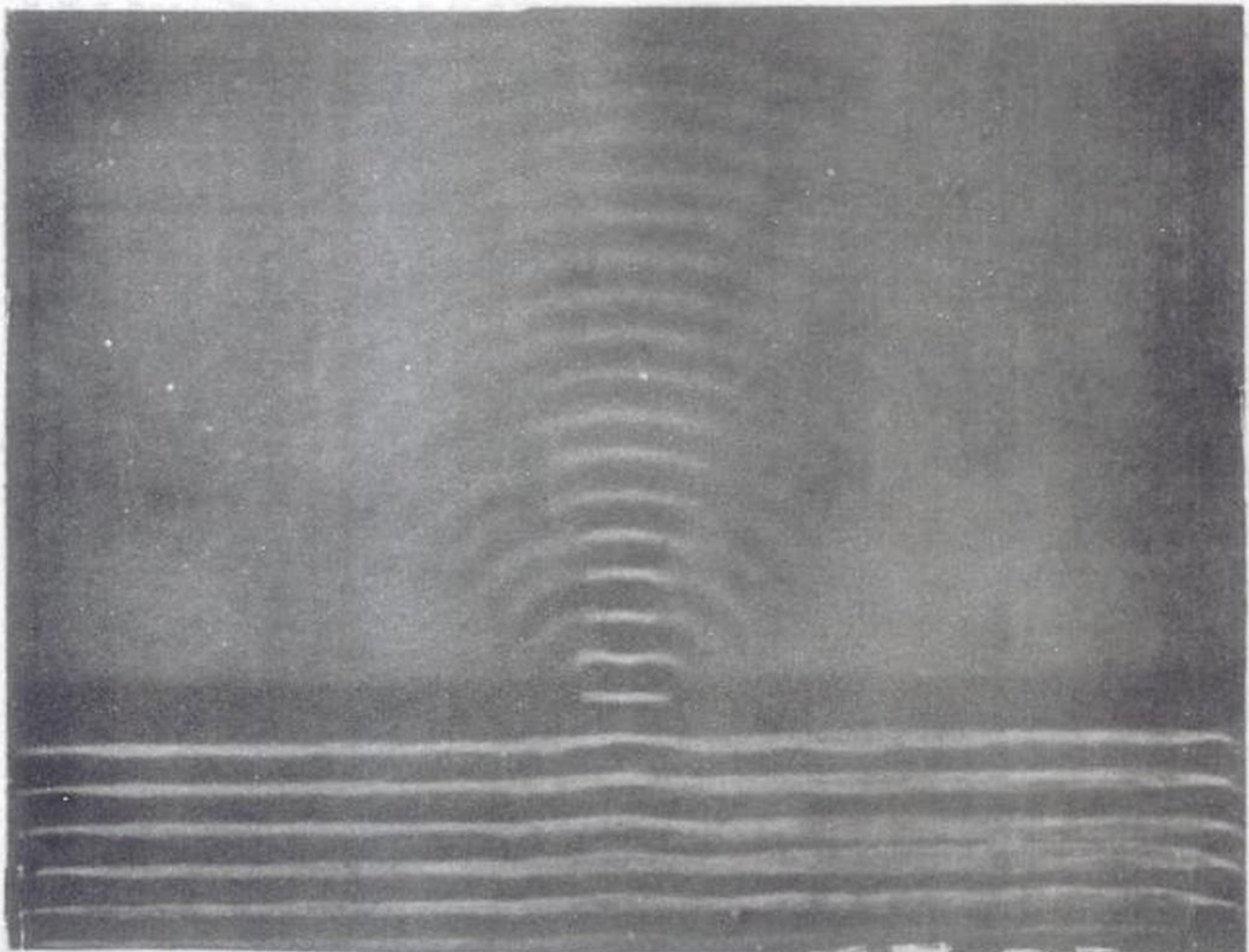


Fig. 10.4 (b) Diffraction of straight water waves when it passes through an opening (adapted from Ref. 6).

*By a tiny hole we imply that the diameter of the hole should be of the order of 0.1 mm or less.

**Use of the Huygens' principle in determining the shape of the wavefront in anisotropic media will be discussed in Chapter 19.

10.4 APPLICATION OF HUYGENS' PRINCIPLE TO STUDY REFRACTION AND REFLECTION

10.4.1 Refraction of a Plane Wave at a Plane Interface

We will first derive the laws of refraction. Let S_1S_2 be a surface separating two media with different speeds of propagation of light v_1 and v_2 as shown in Fig. 10.5. Let A_1B_1 be a plane wavefront incident on the surface at an angle i ; A_1B_1 represents the position of the wavefront at an instant $t = 0$.

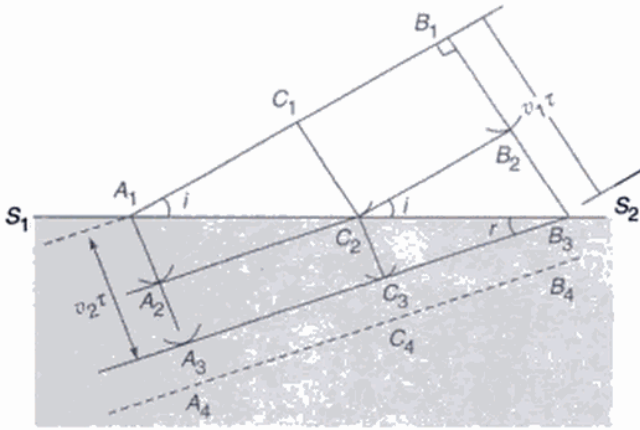


Fig. 10.5 Refraction of a plane wavefront A_1B_1 by a plane interface S_1S_2 separating two media with different velocities of propagation of light v_1 and v_2 ($v_2 < v_1$); i and r are the angles of incidence and refraction respectively. $A_2C_2B_2$ corresponds to the shape of the wavefront at an intermediate time τ_1 . Notice that $r < i$.

Let τ be the time taken for the wavefront to travel the distance B_1B_3 . Then $B_1B_3 = v_1\tau$. In the same time the light would have travelled a distance $A_1A_3 = v_2\tau$ in the second medium. (Note that the lines A_1A_3 , B_1B_3 , etc. are always normal to the wavefront; these represent rays in isotropic media—see Chapter 3). It can easily be seen that the incident and refracted rays make angles i and r with the normal. In order to determine the shape of the wavefront at the instant $t = \tau$ we consider an arbitrary point C_1 on the wavefront. Let the time taken for the disturbance to travel the distance C_1C_2 be τ_1 . Thus $C_1C_2 = v_1\tau_1$. From the point C_2 we draw a secondary wavelet of radius $v_2(\tau - \tau_1)$. Similarly from the point A_1 we draw a secondary wavelet of radius $v_2\tau$. The envelope of these secondary wavelets is

shown as $A_3C_3B_3$. The shape of the wavefront at the intermediate time τ_1 is shown as $A_2C_2B_2$ and clearly $B_1B_2 = C_1C_2 = v_1\tau_1$ and $A_1A_2 = v_2\tau_1$. In the right-angled triangles $B_2C_2B_3$ and $C_3C_2B_3$, $\angle B_2C_2B_3 = i$ (the angle of incidence) and $\angle C_2B_3C_3 = r$ (the angle of refraction). Clearly,

$$\begin{aligned} \frac{\sin i}{\sin r} &= \frac{B_2B_3/C_2B_3}{C_2C_3/C_2B_3} = \frac{B_2B_3}{C_2C_3} \\ &= \frac{v_1(\tau - \tau_1)}{v_2(\tau - \tau_1)} = \frac{v_1}{v_2} \end{aligned} \quad (1)$$

which is known as the Snell's law. It is observed that when light travels from a rarer to a denser medium, the angle of incidence is greater than the angle of refraction and consequently

$$\frac{\sin i}{\sin r} > 1$$

which implies $v_1 > v_2$; thus, Huygens' theory predicts that the speed of light in a rarer medium is greater than the speed of light in a denser medium. This prediction is contradictory to that made by Newton's corpuscular theory (see Sec. 1.2) and as later experiments showed, the prediction of the wave theory was indeed correct.

If c represents the speed of light in free space then the ratio $\frac{c}{v}$ (where v represents the speed of light in the particular medium) is called the refractive index n of the medium. Thus if $n_1 = \left(\frac{c}{v_1}\right)$ and $n_2 = \left(\frac{c}{v_2}\right)$ are the refractive indices of the two media, then Snell's law can also be written as

$$n_1 \sin i = n_2 \sin r \quad (2)$$

Let $A_1C_1B_1$, $A_2C_2B_2$, $A_3C_3B_3$ and $A_4C_4B_4$ denote the successive positions of crests. If λ_1 and λ_2 denote the wavelength of light in medium 1 and medium 2 respectively then, the distance $B_1B_2 (= B_2B_3 = C_1C_2)$ will be equal to λ_1 and the distance $A_1A_2 (= A_2A_3 = C_2C_3)$ will be equal to λ_2 . From Fig. 10.5 it is obvious that

$$\frac{\lambda_1}{\lambda_2} = \frac{\sin i}{\sin r} = \frac{v_1}{v_2} \quad (3)$$

or

$$v_1/\lambda_1 = v_2/\lambda_2 \quad (4)$$

Thus, when a wave gets refracted into a denser medium ($v_1 > v_2$) the wavelength and the speed of propagation decrease but the frequency ($= v/\lambda$) remains the same; when refracted into a rarer medium the wavelength and the speed of propagation will increase. In Table 10.1 we have given the indices of refraction of several materials with respect to vacuum. In Table 10.2, the wavelength dependence of the refractive index for crown glass and vitreous quartz are

Table 10.1 Refractive Indices of various Materials Relative to Vacuum (Adapted from Ref. 1)

 (For light of wavelength $\lambda = 5.890 \times 10^{-5} \text{ cm}$)¹

Material	n	Material	n
Vacuum	1.0000	Quartz (fused)	1.46
Air	1.0003	Rock salt	1.54
Water	1.33	Glass (ordinary crown)	1.52
Quartz (crystalline)	1.54	Glass (dense flint)	1.66

Table 10.2 Refractive Indices of Telescope Crown Glass and Vitreous Quartz for Various Wavelengths (Adapted from Ref. 7)

	Wavelength	Telescope crown	Vitreous quartz
1	$6.562816 \times 10^{-5} \text{ cm}$	1.52441	1.45640
2	$5.889953 \times 10^{-5} \text{ cm}$	1.52704	1.45845
3	$4.861327 \times 10^{-5} \text{ cm}$	1.53303	1.46318

Note: The wavelengths specified at serial numbers 1, 2 and 3 correspond roughly to the red, yellow and blue colours. The table shows the accuracy with which the wavelengths and refractive indices can be measured.

given. The three wavelengths correspond roughly to the red, yellow and blue colours. Notice the accuracy with which the wavelength and the refractive index can be measured.

10.4.2 Total Internal Reflection

In Fig. 10.5 the angle of incidence has been shown to be greater than the angle of refraction. This corresponds to the case when $v_2 < v_1$, i.e., the light wave is incident on a denser medium. However, if the second medium is a rarer medium (i.e. $v_1 < v_2$) then the angle of refraction will be greater than the angle of incidence, and a typical refracted wavefront would be of the form as shown in Fig. 10.6, where $B_1B_2 = v_1\tau$ and $A_1A_2 = v_2\tau$. Clearly, if the angle of incidence is such that $v_2\tau$ is greater than A_1B_2 , then the refracted wavefront will be absent and we will have, what is known as, total internal reflection. The critical angle will correspond to

$$A_1B_2 = v_2\tau$$

Thus

$$\sin i_c = \frac{B_1B_2}{A_1B_2} = \frac{v_1}{v_2} = n_{12}, \quad (5)$$

where i_c denotes the critical angle and n_{12} represents the refractive index of the second medium with respect to the first. For all angles of incidence greater than i_c , we will have total internal reflection.

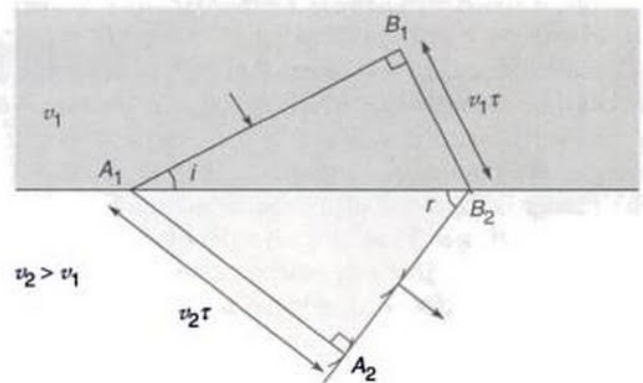


Fig. 10.6 Refraction of a plane wavefront incident on a rarer medium (i.e. $v_2 > v_1$). Notice that the angle of refraction r is greater than the angle of incidence i . The value of i , when r is equal to $\pi/2$, gives the critical angle.

10.4.3 Reflection of a Plane Wave by a Plane Surface

Let us consider a plane wave AB incident at an angle i on a plane mirror as shown in Fig. 10.7. We consider the reflection of the plane wave and try to obtain the shape of the reflected wavefront. Let the position of the wavefront at $t = 0$ be AB . If the mirror was not present, then at a later time τ the position of the wavefront would have been CB' , where $BB' = PP' = AC = v\tau$ and v is the speed of propagation of the wave. In order to determine the shape of the reflected wavefront at the instant $t = \tau$, we consider an arbitrary point P on the wavefront AB and let τ_1 be the time taken by a disturbance to reach the point P_1 from P . From the point P_1 ,

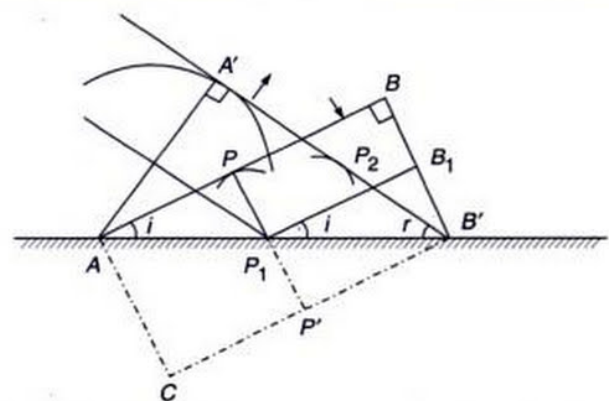


Fig. 10.7 Reflection of a plane wavefront AB incident on a plane mirror. $A'B'$ is the reflected wavefront; i and r correspond to angles of incidence and reflection respectively.

we draw a sphere of radius $v(\tau - \tau_1)$. We draw a tangent plane on this sphere from the point B' . Since $BB_1 = PP_1 = v\tau_1$, the distance B_1B' will be equal to $P_1P_2 [= v(\tau - \tau_1)]$. If we consider triangles P_2P_1B' and B_1P_1B' then the side P_1B' is common to both and since $P_1P' = B'B_2$, and since both the triangles are right-angled triangles, $\angle P_2B'P_1 = \angle B_1P_1B'$. The former is the angle of reflection and the latter is the angle of incidence. Thus, we have the law of reflection; when a plane wavefront gets reflected from a plane surface, the angle of reflection is equal to the angle of incidence and the reflected wave is a plane wave.

10.4.4 Diffuse Reflection

In the above we have considered the reflection of light from a smooth surface. This is known as specular reflection. If the surface is irregular (as shown in Fig. 10.8) we have, what is known as diffuse reflection. The secondary wavelets emanating from the irregular surface travel in many directions and we do not have a well defined reflected wave. Indeed, it can be shown that if the irregularity in the surface is considerably greater than the wavelength, we will have diffuse reflection.



Fig. 10.8 Diffuse reflection of a plane wavefront from a rough surface. It is evident that one does not have a well-defined reflected beam.

10.4.5 Reflection of Light from a Point Source Near a Mirror

Let us consider spherical waves (emanating from a point source P) incident on a plane mirror MM' , as shown in Fig. 10.9. Let ABC denote the shape of the wavefront at time $t = 0$. In the absence of the mirror, the shape of the wavefront at a later time τ would have been $A_1B_1C_1$ where $AA_1 = BB_1 = CC_1 = v\tau$, Q being an arbitrary point on the wavefront. If the time taken for the disturbance to traverse the distance QQ' be τ_1 then, in order to determine the shape of the reflected wavefront, we draw a sphere of radius $v(\tau - \tau_1)$ whose center is at the point Q' . In a similar manner we can draw the secondary wavelets emanating from other points on the mirror and, in particular, from the point B we have to draw a sphere of radius $v\tau$. The shape of the reflected wavefront is obtained by drawing a common tangent plane to all these spheres, which is shown as $A_1B_1'C_1$ in the figure. It can immediately be seen that $A_1B_1'C_1$ will

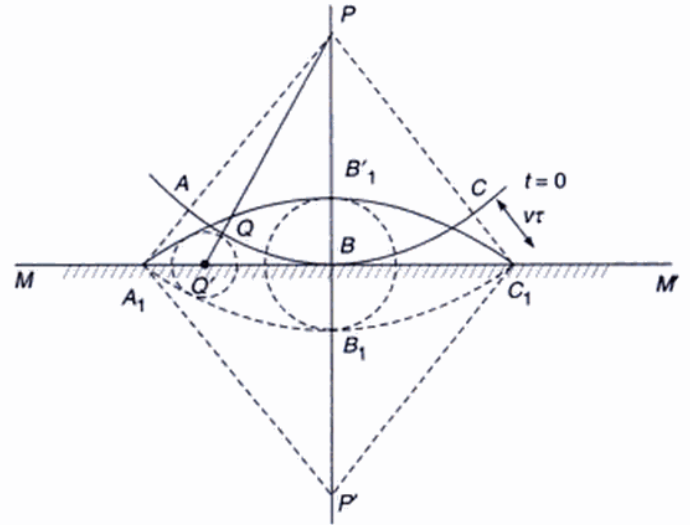


Fig. 10.9 P is a point source placed in front of a plane mirror MM' . ABC is the incident wavefront (which is spherical and centred at P) and $A_1B_1'C_1$ is the corresponding reflected wavefront (which is spherical and centred at P'). P' is the virtual image of P .

have an exactly similar shape as $A_1B_1C_1$ except that $A_1B_1'C_1$ will have its center of curvature at the point P' where $PB = BP'$. Thus the reflected waves will appear to emanate from the point P' which will be the virtual image of the point P .

10.4.6 Refraction of a Spherical Wave by a Spherical Surface

Let us consider spherical waves (emanating from the point P) incident on the curved spherical surface SBS' . Let the shape of the wavefront at the time $t = 0$ be ABC [see Fig. 10.10(a)]. Let the refractive indices on the left and on the right of the spherical surface be n_1 and n_2 respectively. In the absence of the spherical surface, the shape of the wavefront at a later time τ would have been $A_1B_1C_1$ where $AA_1 = BB_1 = CC_1 = v_1\tau$. We consider an arbitrary point Q on the wavefront ABC and let τ_1 be the time taken for the disturbance to reach the point Q' (on the surface of the spherical wave); thus $QQ' = v_1\tau_1$. In order to determine the shape of the refracted wavefront at a later time τ , we draw a sphere of radius $v_2(\tau - \tau_1)$ from the point Q' . We may draw similar spheres from other points on the spherical surface; in particular, the radius of the spherical wavefront from the point B , which is equal to BB_2 will be $v_2\tau$. The envelope of these spherical wavelets is shown as $A_1B_2C_1$ which, in general, will not be a sphere.* However, a small portion of any curved surface can be considered as a sphere and in this

*The fact that the refracted wavefront is not, in general, a sphere leads to, what are known as aberrations.

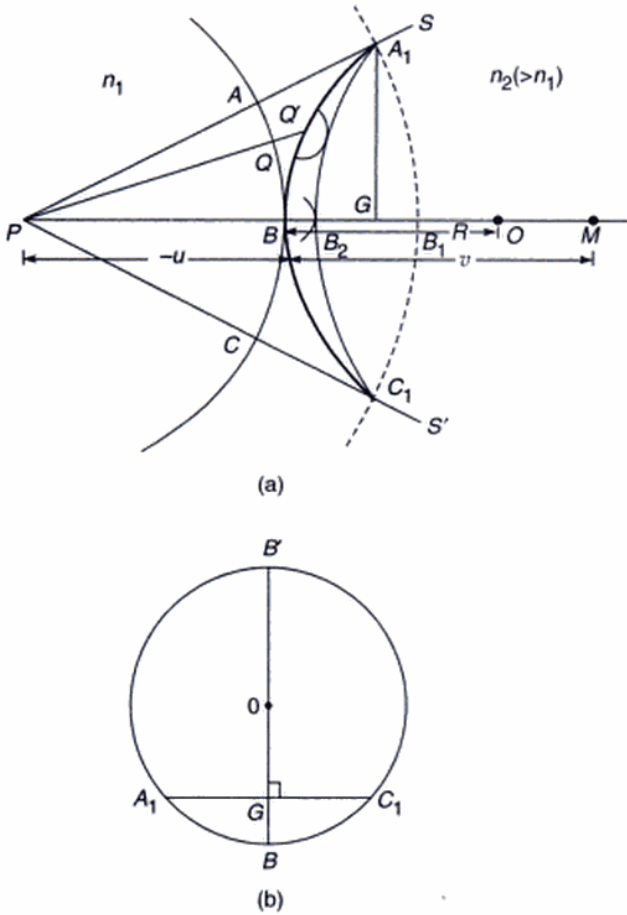


Fig. 10.10 (a) Refraction of a spherical wave ABC (emanating from the point source P) by a convex spherical surface SBS' separating media of refractive indices n_1 and $n_2 (>n_1)$. $A_1B_2C_1$ is the refracted wavefront, which is approximately spherical and whose center of curvature is at M . Thus M is the real image of P . O is the center of curvature of SS' . (b) The diameter $B'O$ intersects the chord A_1GC_1 normally.

approximation we may consider $A_1B_2C_1$ to be a sphere whose center of curvature is at the point M . The spherical wavefront will, therefore, converge towards the point M and hence the point M represents the real image of the point P .

We adopt a sign convention in which all distances, measured to the left of the point B , are negative and all distances measured to the right of the point B are positive. Thus

$$PB = -u$$

where u itself is a negative quantity. Further, since the point M lies on the right of B , we have

$$BM = v$$

and similarly,

$$BO = R$$

where O represents the center of curvature of the spherical surface.

In order to derive a relation between u , v and R we use a theorem in geometry, according to which,

$$(A_1G)^2 = GB \times (2R - GB) \quad (6)$$

where G is the foot of the perpendicular on the axis PM [see Fig. 10.10(b)]. In Fig. 10.10(b) the diameter $B'O$ intersects the chord A_1GC_1 normally. If $GB \ll R$, then

$$(A_1G)^2 \approx 2R(GB)$$

Consider the spherical surface SBS' [see Fig. 10.10(a)] whose radius is R . Clearly,

$$\begin{aligned} (A_1G)^2 &= (2R - GB)GB \\ &\approx 2R(GB) \end{aligned} \quad (7)$$

where we have assumed $GB \ll R$. Similarly by considering the spherical surface $A_1B_2C_1$ (whose center is at the point M) we obtain

$$(A_1G)^2 \approx 2v(GB_2) \quad (8)$$

where $v = BM \approx B_2M$. In a similar manner,

$$(A_1G)^2 \approx 2(-u)GB_1 \quad (9)$$

Since u is a negative quantity, $(A_1G)^2$ is positive. Now

$$BB_1 = v_1\tau \quad \text{and} \quad BB_2 = v_2\tau$$

Therefore

$$\frac{BB_1}{BB_2} = \frac{v_1}{v_2} = \frac{n_2}{n_1}$$

or

$$n_1BB_1 = n_2BB_2$$

or

$$n_1(BG + GB_1) = n_2(BG - GB_2)$$

or

$$n_1 \left[\frac{(A_1G)^2}{2R} - \frac{(A_1G)^2}{2u} \right] = n_2 \left[\frac{(A_1G)^2}{2R} - \frac{(A_1G)^2}{2v} \right]$$

where we have used Eqs (7), (8) and (9). Thus

$$\frac{n_2}{v} - \frac{n_1}{u} = \frac{n_2 - n_1}{R} \quad (10)$$

which may be rewritten in the form

$$\frac{n_2}{v} = \frac{n_1}{u} + \frac{n_2 - n_1}{R} \quad (11)$$

Thus, if

$$\frac{n_1}{|u|} > \frac{n_2 - n_1}{R}$$

or

$$|u| < \frac{Rn_1}{n_2 - n_1}$$

we will obtain a virtual image. (We are of course assuming that the second medium is a denser medium, i.e., $n_2 > n_1$; if $n_2 < n_1$, we will always have a virtual image).

A converging spherical wavefront will propagate in a manner shown in Fig. 10.11. Beyond the focal point it will start diverging as shown in the figure.*

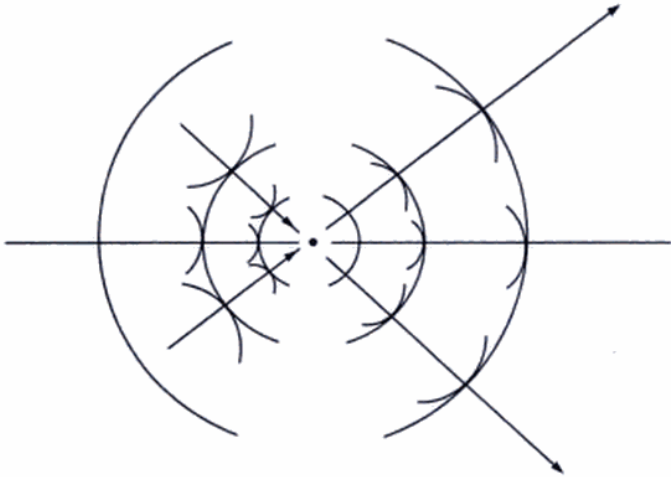


Fig. 10.11 Propagation of a converging spherical wave using Huygens' principle.

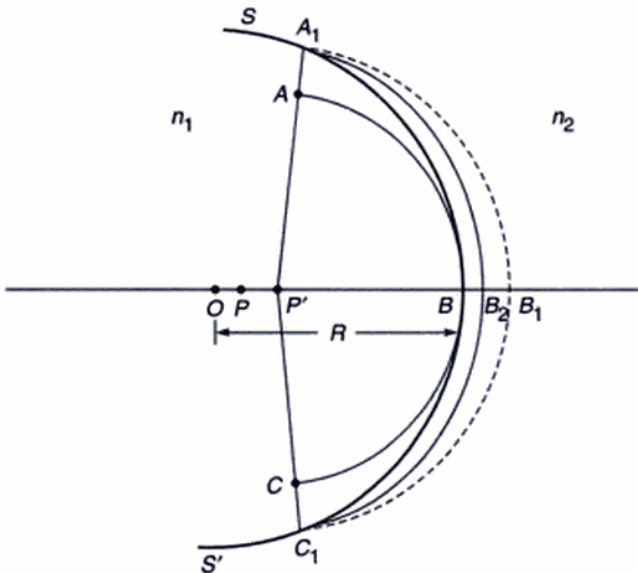


Fig. 10.12 Refraction of a spherical wave by a concave surface separating media of refractive indices n_1 and n_2 ($> n_1$). P' is the virtual image of P .

In a similar manner we can consider the refraction of a spherical wave from a surface SBS' shown in Fig. 10.12 ($n_2 > n_1$). Here the center of curvature will also lie on the left of the point B and both u and R will be negative quantities. Thus no matter what the values of u and R may be, v will be negative and we will obtain a virtual image.

Using Eq. (10) we can easily derive the thin lens formula. We assume a thin lens made of a material of refractive index n_2 to be placed in a medium of refractive index n_1 (See Fig. 10.13). Let the radii of curvatures of the first and the second surface be R_1 and R_2 respectively. Let v' be the distance of the image of the object P if the second surface were not present. Then

$$\frac{n_2}{v'} - \frac{n_1}{u} = \frac{n_2 - n_1}{R_1} \quad (12)$$

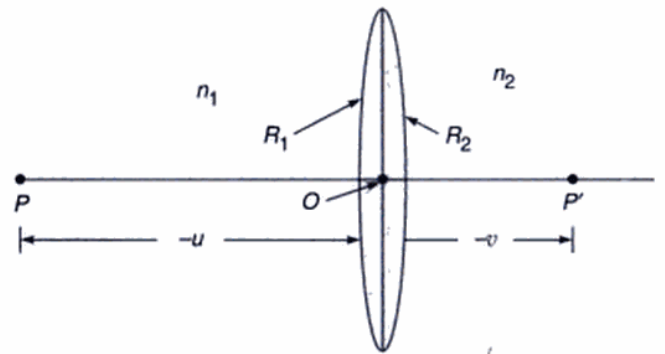


Fig. 10.13 A thin lens made of a medium of refractive index n_2 placed in a medium of refractive index n_1 . The radii of curvatures of the two surfaces are R_1 and R_2 . P is the image (at a distance v from the point O) of the point object P (at a distance $-u$ from the point O).

(Since the lens is assumed to be thin, all the distances are measured from the point O). This image now acts as an object to the spherical surface R_2 on the left of which is the medium of refractive index n_2 and on the right of which is the medium of refractive index n_1 . Thus, if v is the distance of the final image point from O , then

$$\frac{n_1}{v} - \frac{n_2}{v'} = \frac{n_1 - n_2}{R_2} \quad (13)$$

Adding Eqs (12) and (13), we obtain

$$\frac{n_1}{v} - \frac{n_1}{u} = (n_2 - n_1) \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (14)$$

or,

$$\frac{1}{v} - \frac{1}{u} = \frac{1}{f} \quad (15)$$

* Very close to the focal point, one has to use a more rigorous wave theory and the shape of the wavefront is very much different from spherical (see Ref. 8). However, much beyond the focal point the wavefronts again become spherical.

where

$$\frac{1}{f} = \frac{n_2 - n_1}{n_1} \left(\frac{1}{R_1} - \frac{1}{R_2} \right) \quad (16)$$

Notice that we do not have to worry whether v' is positive or negative; it is automatically taken care of through the sign convention. Further, the relation derived is valid for any lens; for example, for a double convex lens, R_1 is positive and R_2 is negative and for a double concave lens, R_1 is negative and R_2 is positive. Similarly it follows for other types of lenses (see Fig. 4.6).

Example 10.1 Consider a vibrating source moving through a medium with a speed V . Let the speed of propagation of the wave in the medium be v . Show that if $V > v$ then a conical wavefront is set up whose half-angle is given by

$$\theta = \sin^{-1} \left(\frac{v}{V} \right) \quad (17)$$

Solution: Let at $t = 0$, the source be at the point P_0 moving with a speed V in the x -direction (see Fig. 10.14). We wish to find out the wavefront at a later time τ . The disturbance emanating from the point P_0 traverses a distance $v\tau$ in time τ . Thus from the point P_0 we draw a sphere of radius $v\tau$. We next consider the waves emanating from the source at a time τ_1 ($< \tau$). At time τ_1 let the source be at the position P_1 ; consequently,

$$P_0P_1 = V\tau_1$$

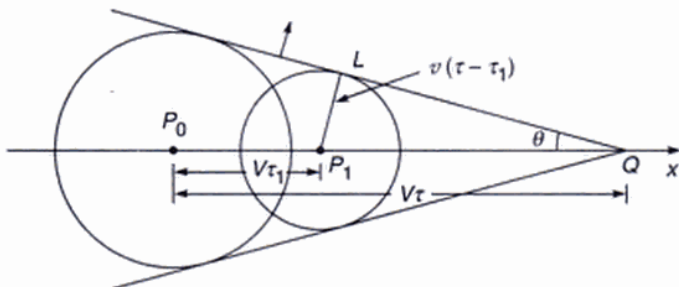


Fig. 10.14 Generation of a shock wavefront by a vibrating particle P_0 moving with a speed V , in a medium in which the velocity of propagation of the wave is v ($< V$).

In order to determine the shape of the wavefront at τ , we draw a sphere of radius $v(\tau - \tau_1)$ centered at P_1 . Let the source be at the position Q at the instant τ . Then

$$P_0Q = V\tau$$

We draw a tangent plane from the point Q , on the sphere whose origin is the point P_1 . Since

$$P_1L = v(\tau - \tau_1) \quad \text{and} \quad P_1Q = V(\tau - \tau_1)$$

$$\sin \theta = \frac{P_1L}{P_1Q} = \frac{v}{V} \quad (\text{independent of } \tau_1)$$

Since θ is independent of τ_1 , all the spheres drawn from any point on the line P_0Q will have a common tangent plane. This plane is known as the shock wavefront and propagates with a speed v .

It is interesting to point out that even when the source is not vibrating, if its speed is greater than the speed of sound waves, a shock wavefront is always set up. A similar phenomenon also occurs when a charged particle (like an electron) moves in a medium with a speed greater than the speed of light in that medium.* The emitted light is known as Cerenkov radiation. If you ever see a swimming pool type reactor, you will find a blue glow coming out from it; this is because of the Cerenkov radiation emitted by the fast moving electrons.

10.5 HUYGENS' PRINCIPLE IN INHOMOGENEOUS MEDIA

Huygens' principle can also be used to study the propagation of a wavefront in an inhomogeneous medium. For definiteness, we consider a medium whose refractive index decreases continuously from a given axis, which we define as the z -axis; x and y -axis being the transverse axis. A simple example is a Selfoc fiber,** whose refractive index variation is of the form

$$n^2(x, y) = n_1^2 - \gamma^2(x^2 + y^2) \quad (18)$$

where n_1 is the refractive index on the z -axis. Let the plane wavefront be incident along the z -axis as shown in Fig. 10.15. Since the refractive index decreases as x and y increase, the speed of the secondary wavelets emanating from portions of the incident wavefront will increase as we move away from the axis. Let us try to determine the shape of the wavefront at a time Δt ; given that the wavefront at $t = 0$ is a plane wavefront A_1B_1 (see Fig. 10.15). We will have to draw spheres of radius $v(x, y)\Delta t$, centered at (x, y) , where $v(x, y)$ is the velocity of the wave at the point (x, y) , which increases as x and y increase. Thus the radii of the spheres increase as we move away from the axis and if we draw a

*This does not contradict the theory of relativity according to which no particle can have a speed greater than the speed of light in free space ($= 3 \times 10^8$ m/sec). The speed of light in a medium will be equal to c/n , where n represents the refractive index. For example in water, the speed of light will be about 2.25×10^8 m/sec and the speed of the electron could be greater than this value.

**See Sec. 2.4.1

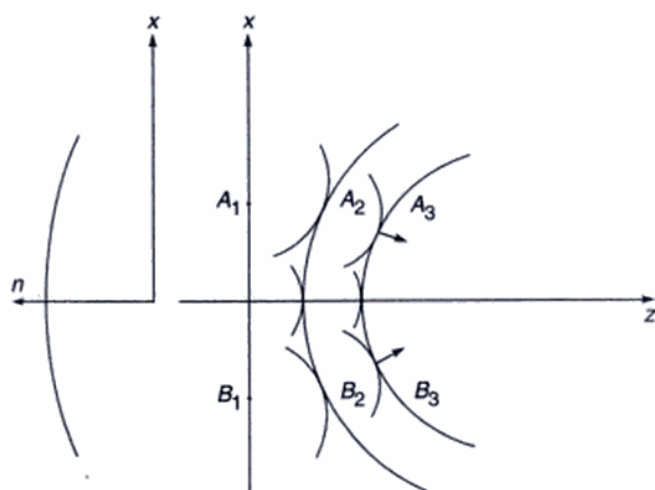


Fig. 10.15 The focusing of an incident plane wavefront in an inhomogeneous medium characterized by a refractive index variation given by Eq. (18).

common tangent to all these spheres then the resulting wavefront is shown in Fig. 10.15 as A_2B_2 . It is at once evident that the wavefront which was initially plane has now become curved. If we again use the same procedure, then the shape of the wavefront at time $2\Delta t$ (say) is shown as A_3B_3 . Thus it is evident that in the present case the wavefront is getting focused. It should be borne in mind that since we are considering an inhomogeneous medium, the refractive index varies continuously with position. For the above construction to be valid, Δt should be small so that during this short interval the secondary wavelets may be assumed to be spherical.

SUMMARY

- According to Huygens' principle, each point of a wavefront is a source of secondary disturbance and the wavelets emanating from these points spread out in all directions with the speed of the wave. The envelope of these wavelets gives the shape of the new wavefront.
- Huygens' principle along with the fact that the secondary wavelets mutually interfere, is known as the Huygens – Fresnel principle.
- Laws of reflection and Snell's law of refraction can be derived using Huygens' principle.
- Using Huygens' principle one can derive the lens

$$\text{formula } \frac{1}{v} - \frac{1}{u} = \frac{1}{f}.$$

PROBLEMS

- 10.1** Use Huygens' principle to study the reflection of a spherical wave emanating from a point on the axis at a concave mirror of radius of curvature R and obtain the mirror equation

$$\frac{1}{u} + \frac{1}{v} = \frac{2}{R}$$

- 10.2** Consider a plane wave incident obliquely on the face of a prism. Using Huygens' principle, construct the transmitted wavefront and show that the deviation produced by the prism is given by

$$\delta = i + t - A$$

where A is the angle of the prism, i and t are the angles of incidence and transmittance.

REFERENCES AND SUGGESTED READINGS

1. A.B. Arons, *Development of Concepts in Physics*, Addison-Wesley Publishing Co., Reading, Mass., 1965.
2. B.B. Baker and E.J. Copson, *The Mathematical Theory of Huygens' Principle*, Oxford University Press, London, 1969.
3. H.J.J. Braddick, *Vibration, Waves and Diffraction*, McGraw-Hill Publishing Co., London, 1965.
4. A.J DeWitte, 'Equivalence of Huygens' principle and Fermat's principle in ray geometry', *America Journal of Physics*, Vol. 27, 293, 1959.
5. C. Huygens, *Treatise on Light*, Dover Publications, New York, 1962.
6. PSSC, *Physics*, D.C. Heath and Company, Boston, Mass., 1965.
7. F.A. Jenkins and H.E. White, *Fundamentals of Optics*, 3rd Edn., McGraw-Hill, 1957, p. 465.
8. M. Born and E. Wolf, *Principle of Optics*, Pergamon Press, Oxford, 1975.

PART 3

Interference

This part commences with the principle of superposition of waves which is the basic physics behind all interference experiments. Starting with the Young's double hole interference experiment, the various chapters discuss many interesting experiments associated with the area of interference. In Chapter 13 the Michelson Interferometer is discussed in detail which is perhaps one of the most ingenious and sensational optical instrument for which Michelson received the Nobel Prize in Physics in 1907. Chapter 14 discusses the Fabry – Perot interferometer which is based on multiple beam interference and are characterized by a high resolving power and hence find applications in high resolution spectroscopy. Chapter 15 discusses the basic concept of temporal and spatial coherence. The ingenious experiment of Michelson (which used the concept of spatial coherence to determine the angular diameter of stars), optical beats and Fourier transform spectroscopy have also been discussed.

Chapter 11

Superposition of Waves

The experiments described appear to me, at any rate, eminently adapted to remove any doubt as to the identity of light, radiant heat, and electromagnetic wave motion. I believe that from now on we shall have greater confidence in making use of the advantages which this identity enables us to derive both in the study of optics and of electricity.

— Heinrich Hertz (1888)*

11.1 INTRODUCTION

In this chapter we will discuss the applications of the principle of superposition of waves according to which the resultant displacement (at a particular point) produced by a number of waves is the vector sum of the displacements produced by each one of the disturbances. As a simple example, we consider a long stretched string AB (see Fig. 11.1). From the end A , a triangular pulse is generated which propagates to the right with a certain speed v . In the absence of any other disturbance, this pulse would have propagated in the $+x$ -direction without any change in shape; we are, of course, neglecting any attenuation or distortion of the pulse. We next assume that from the end B an identical pulse is generated which starts moving to the left with the same speed v . (As has been shown in Sec. 9.6 the speed of the wave is determined by the ratio of the tension in the string to its mass per unit length.) At $t = 0$, the snapshot of the string is shown in Fig. 11.1(a). At a little later time each pulse moves close to the other as shown in Fig. 11.1(b), without any interference. Figure 11.1(c) represents a snapshot at an instant when the two pulses interfere; the dashed curves represent the profile of the string if each of the impulses was moving all by itself, whereas the solid curve shows the resultant displacement obtained by algebraic addition of each displacement. Shortly later [Fig. 11.1(d)] the two pulses exactly overlap each other and the resultant displacement is zero everywhere (where has the energy gone?). At a much later time the impulses sort of cross each other [Fig. 11.1(e)] and move as if nothing had happened. This is a characteristic feature of superposition of waves.

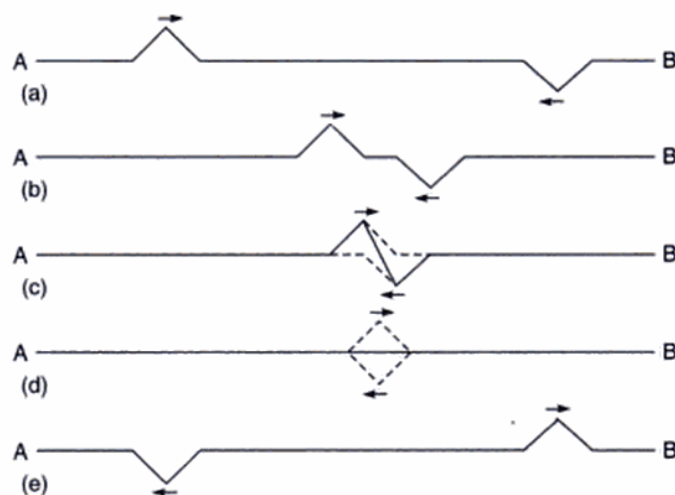


Fig. 11.1 The propagation in opposite directions of two triangular pulses in a stretched string. The solid line gives the actual shape of the string; (a), (b), (c), (d) and (e) correspond to different instants of time.

The phenomenon of interference contains no more physics than embodied in the above example. In the following sections we will consider some more examples.

11.2 STATIONARY WAVES ON A STRING

Consider a string which is fixed at the point A (see Fig. 11.2). A transverse sinusoidal wave is sent down the string along the $-x$ -direction. The displacement at any

*The author found this quotation in the book by Smith and King (Ref. 1).

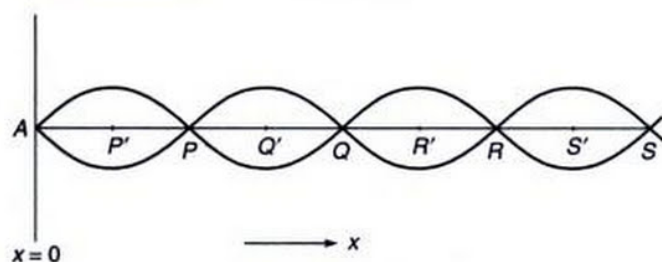


Fig. 11.2 Reflection of a wave at $x = 0$.

point on the string due to this wave would be given by

$$y_i = a \sin \left[\frac{2\pi}{\lambda} (x + vt) + \phi \right] \quad (1)$$

where the subscript i refers to the fact that we are considering the incident wave. Without any loss of generality we can set $\phi = 0$; thus we may write

$$\begin{aligned} y_i &= a \sin \left[\frac{2\pi}{\lambda} (x + vt) \right] \\ &= a \sin \left[2\pi \left(\frac{x}{\lambda} + vt \right) \right] \end{aligned} \quad (2)$$

Thus, because of the incident wave, the displacement at the point A would have been

$$y_i \big|_{x=0} = a \sin (2\pi vt) \quad (3)$$

where $v = v/\lambda$ and we have assumed the point A to correspond to $x = 0$. Since the point A is fixed, there must be a reflected wave such that the displacement due to this reflected wave (at the point A) is equal and opposite to y_i :

$$y_r \big|_{x=0} = -a \sin (2\pi vt) \quad (4)$$

where the subscript r refers to the fact that we are considering the reflected wave. Since the reflected wave propagates in the $+x$ -direction we must have

$$y_r = +a \sin 2\pi \left(\frac{x}{\lambda} - vt \right) \quad (5)$$

The resultant displacement would be given by

$$\begin{aligned} y &= y_i + y_r = a \left[\sin 2\pi \left(\frac{x}{\lambda} + vt \right) + \sin 2\pi \left(\frac{x}{\lambda} - vt \right) \right] \\ &= 2a \sin \frac{2\pi}{\lambda} x \cos 2\pi vt \end{aligned} \quad (6)$$

It should be seen that for values of x such that

$$\sin \frac{2\pi}{\lambda} x = 0 \quad (7)$$

the displacement y is zero at all times. Such points are known as *nodes*; the x -coordinates of the nodes are given by

$$x = 0, \frac{\lambda}{2}, \lambda, \frac{3\lambda}{2}, 2\lambda, \dots \quad (8)$$

and are marked as points A, P, Q and R in Fig. 11.2. The nodes are separated by a distance $\lambda/2$ and at the midpoint between two consecutive nodes, i.e. at

$$x = \frac{\lambda}{4}, \frac{3\lambda}{4}, \frac{5\lambda}{4}, \dots$$

the amplitude of the vibration is maximum. The displacements at these points (which are known as *antinodes*) are given by

$$y = \pm 2a \cos 2\pi vt \quad (9)$$

At the antinodes the kinetic energy density would be given by (see Sec. 6.2)

$$\begin{aligned} \text{Kinetic energy/unit length} &= \frac{1}{2} \rho (2a)^2 \omega^2 \cos^2 \omega t \\ &= 2\rho a^2 \omega^2 \cos^2 \omega t \end{aligned} \quad (10)$$

where $\omega = 2\pi v$ is the angular frequency and ρ the mass per unit length of the string.

We can also carry out a similar experiment for electromagnetic waves. In Fig. 11.3, T represents a transmitter of electromagnetic waves (the wavelength of which may be of

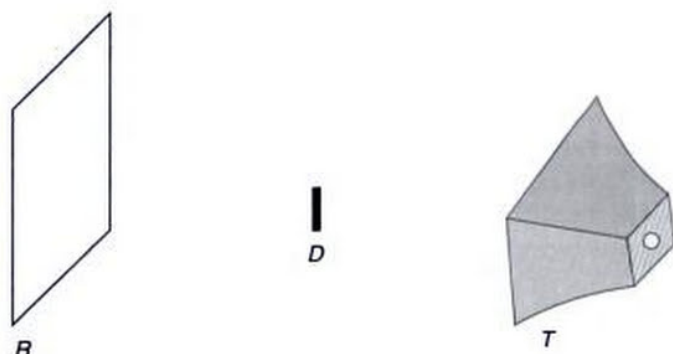


Fig. 11.3 An arrangement for studying standing electromagnetic waves.

the order of few cm); R represents a reflector which may be a highly polished metal surface and D represents the detector which can measure the variation of the intensity of the electromagnetic waves at different points. One may approximately assume plane waves to be incident on the reflector; the incident and reflected waves interfere and produce nodes and antinodes. The result of a typical experiment is shown in Fig. 11.4. One can see the periodic variation of intensity. Two consecutive maxima are separated by about 5.8 cm; thus $\lambda \approx 11.6$ cm. The corresponding frequency ($\approx 2.6 \times 10^9 \text{ sec}^{-1}$) can easily be generated in the laboratory. If the frequency is changed, one can observe the

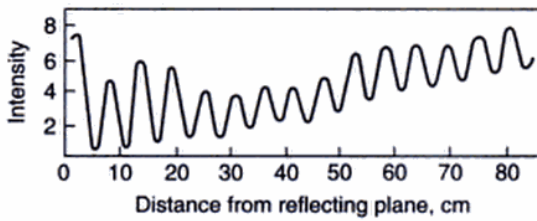


Fig. 11.4 A typical variation of the intensity between the reflector and the transmitter [adapted from Ref. 2].

change in the distance between the antinodes. One should notice that the minima do not really correspond to zero intensity and the intensities at the maxima are not constant. This is because of the fact that the incident wave is really not a plane wave* and that the reflection is not really perfect. In fact, one can introduce a coefficient of reflection (r) which is defined as the ratio of the energy of the reflected beam to the energy of the incident beam. Thus the ratio of the amplitudes would be \sqrt{r} and if the incident wave is given by

$$E_{\text{incident}} = a \sin \left[2\pi \left(\frac{x}{\lambda} + vt \right) \right] \quad (11)$$

then the reflected wave would be given by

$$E_{\text{reflected}} = a\sqrt{r} \sin \left[2\pi \left(\frac{x}{\lambda} - vt \right) \right] \quad (12)$$

where the plane $x = 0$ corresponds to the plane of the reflector. Here E represents the electric field associated with the electromagnetic wave. Thus the resultant field would be given by

$$\begin{aligned} E_{\text{resultant}} &= E_{\text{incident}} + E_{\text{reflected}} \\ &= a \sin \left[2\pi \left(\frac{x}{\lambda} + vt \right) \right] + a\sqrt{r} \sin \left[2\pi \left(\frac{x}{\lambda} - vt \right) \right] \\ &= a\sqrt{r} \left\{ \sin \left[2\pi \left(\frac{x}{\lambda} + vt \right) \right] + \sin \left[2\pi \left(\frac{x}{\lambda} - vt \right) \right] \right\} \\ &\quad + a(1 - \sqrt{r}) \sin \left[2\pi \left(\frac{x}{\lambda} + vt \right) \right] \\ &= 2a\sqrt{r} \sin \left(\frac{2\pi}{\lambda} x \right) \cos 2\pi vt \\ &\quad + a(1 - \sqrt{r}) \sin \left[2\pi \left(\frac{x}{\lambda} + vt \right) \right] \end{aligned}$$

The first term represents the stationary component of the wave and the second term (which is small if r is close to unity) represents the progressive part of the beam.

*A plane wave is obtained by a point source at a very large distance from the point of observation (see Chapter 9).

11.3 STATIONARY WAVES ON A STRING WHOSE ENDS ARE FIXED

In Sec. 11.2, while discussing the stationary waves on a string we had assumed only one end of the string ($x = 0$) to be fixed; and the resultant displacement was shown to be given by [see Eq. (6)]:

$$y = 2a \sin \left(\frac{2\pi}{\lambda} x \right) \cos (2\pi vt) \quad (14)$$

If the other end of the string (say at $x = L$) is also fixed, then we must have

$$2a \sin \left(\frac{2\pi}{\lambda} L \right) \cos (2\pi vt) = 0 \quad (15)$$

Eq. (15) is to be valid at all times, therefore,

$$\sin \left(\frac{2\pi}{\lambda} L \right) = 0 = \sin n\pi \quad (16)$$

or

$$\lambda = \lambda_n = \frac{2L}{n}, \quad n = 1, 2, 3, \dots \quad (17)$$

The corresponding frequencies are

$$\nu_n = \frac{v}{\lambda_n} = \frac{n\nu}{2L}, \quad n = 1, 2, 3, \dots \quad (18)$$

Thus, if a string of length L is clamped at both ends (as in a sonometer wire) then it can only vibrate with certain well defined wavelengths. When $\lambda = 2L$ (i.e., $n = 1$) the string is said to vibrate in its fundamental mode [Fig. 11.5(a)]. Similarly when $\lambda = 2L/2$ and $2L/3$ the string is said to vibrate in its first and second harmonic. In general,

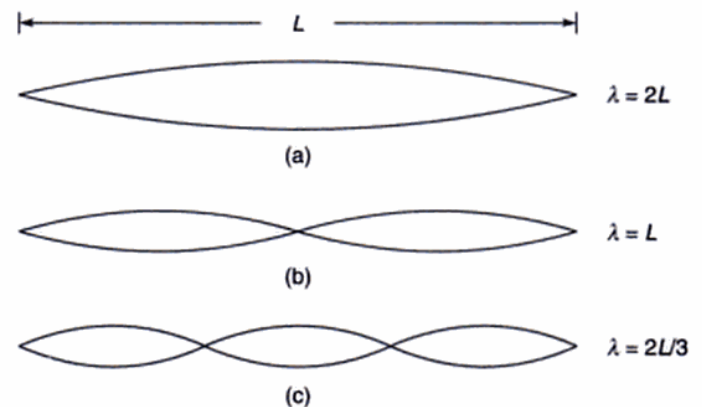


Fig. 11.5 Standing waves on a stretched string clamped at both ends.

if the string is plucked and then made to vibrate then the displacement would be given by

$$y(x, t) = \sum_{n=1}^{\infty} a_n \sin\left(\frac{2\pi}{\lambda_n} x\right) \cos(2\pi \nu_n t + \phi_n) \quad (19)$$

where the constants a_n and ϕ_n are determined by the values of $y(x, t = 0)$ and $\left.\frac{\partial y}{\partial t}\right|_{t=0}$; these are known as the initial

conditions. A more detailed discussion on the vibration of stretched strings has been given in Sec. 7.2.

It should be seen that when a string is vibrating in a particular mode there is no net transfer of energy although each element of the string is associated with a certain energy density [see Eq. (10)]. The energy density is maximum at the antinodes and minimum at nodes. The distances between two successive antinodes and successive nodes are $\lambda/2$.

11.4 STATIONARY LIGHT WAVES: IVES AND WIENER'S EXPERIMENTS

It is difficult to carry out experiments in which one obtains stationary light waves. This is because of the fact that light wavelengths are extremely small ($\approx 5 \times 10^{-5}$ cm). In the experimental arrangement of Ives, the emulsion side of a photographic plate was placed in contact with a film of mercury as shown in Fig. 11.6. A parallel beam of monochromatic light was allowed to fall normally on the glass plate. The beam was reflected on the mercury surface and

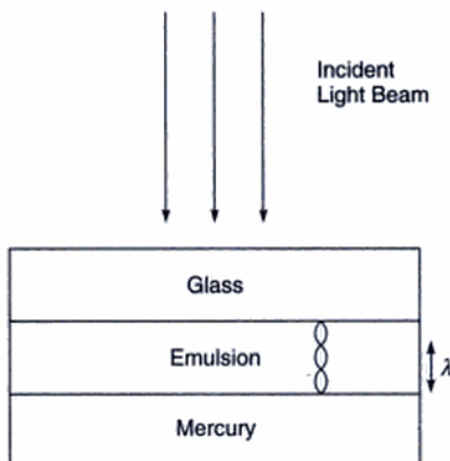


Fig. 11.6 The experimental arrangement of Ives for studying stationary light waves.

the incident wave interfered with the reflected wave forming standing waves. A section of the photographic film was cut along a plane normal to the surface. The cut section was viewed under a microscope and bright and dark bands (separated by regular intervals) were observed. By measuring the distance between two consecutive dark bands (which is equal to $\lambda/2$) one can calculate the wavelength.

Because of the small wavelength of light, the distance between two consecutive dark (or bright) bands was extremely small and was, therefore, difficult to measure. Wiener overcame this difficulty by placing the photographic film at a small angle and thereby increasing considerably the distance between the dark (or bright) bands (Fig. 11.7).

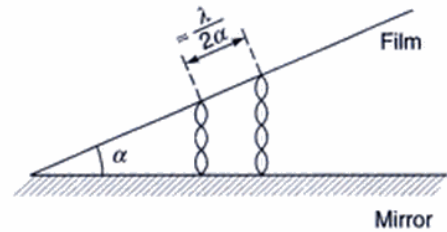


Fig. 11.7 The experimental arrangement of Wiener for studying stationary light waves.

Example 11.1 In a typical experimental arrangement of Wiener, the angle between the film and the mirror was about 10^{-3} radians. For $\lambda = 5 \times 10^{-5}$ cm what would be the distance between two consecutive dark bands?

Solution: The required distance is

$$\frac{\lambda}{2\alpha} = \frac{5 \times 10^{-5}}{2 \times 10^{-3}} \text{ cm} = 0.25 \text{ mm}$$

On the other hand, in the set up of Ives the distance would be 2.5×10^{-4} mm.

11.5 SUPERPOSITION OF TWO SINUSOIDAL WAVES

Let us consider the superposition of two sinusoidal waves (having the same frequency*) at a particular point. Let

$$\text{and } \left. \begin{aligned} x_1(t) &= a_1 \cos(\omega t + \theta_1) \\ x_2(t) &= a_2 \cos(\omega t + \theta_2) \end{aligned} \right\} \quad (20)$$

represent the displacements produced by each of the disturbances: we are assuming that the displacements are in

*In Chapter 15 we will consider the superposition of waves having nearly equal frequencies which leads to the phenomenon of beats.

the same direction*, however, they may have different amplitudes and different initial phases. Now, according to the superposition principle the resultant displacement $x(t)$ would be given by

$$x(t) = x_1(t) + x_2(t) \\ = a_1 \cos(\omega t + \theta_1) + a_2 \cos(\omega t + \theta_2) \quad (21)$$

which can be written in the form

$$x(t) = a \cos(\omega t + \theta) \quad (22)$$

where

$$a \cos \theta = a_1 \cos \theta_1 + a_2 \cos \theta_2 \quad (23)$$

and

$$a \sin \theta = a_1 \sin \theta_1 + a_2 \sin \theta_2 \quad (24)$$

Thus the resultant disturbance is also simple harmonic in character having the same frequency but different amplitude and different initial phase. If we square and add Eqs (23) and (24), we would obtain

$$a = [a_1^2 + a_2^2 + 2a_1a_2 \cos(\theta_1 - \theta_2)]^{1/2} \quad (25)$$

Further

$$\tan \theta = \frac{a_1 \sin \theta_1 + a_2 \sin \theta_2}{a_1 \cos \theta_1 + a_2 \cos \theta_2} \quad (26)$$

It should be pointed out that θ is not uniquely determined from Eq. (26); however, if we assume a to be always positive, then $\cos \theta$ and $\sin \theta$ can be determined from Eqs (23) and (24) which will uniquely determine θ .

From Eq. (25) we find that if

$$\theta_1 \sim \theta_2 = 0, 2\pi, 4\pi, \dots \quad (27)$$

$$\text{then } a = a_1 + a_2 \quad (28)$$

Thus, if the two displacements are in phase, then the resultant amplitude will be the sum of the two amplitudes; this is known as **constructive interference**. Similarly, if

$$\theta_1 \sim \theta_2 = \pi, 3\pi, 5\pi, \dots \quad (29)$$

then

$$a = a_1 - a_2 \quad (30)$$

and the resultant amplitude is the difference of the two amplitudes. This is known as **destructive interference**. If we refer to Fig. 11.2, then we can see that constructive interference occurs at $x = \frac{\lambda}{4}, \frac{3\lambda}{4}, \frac{5\lambda}{4}, \dots$ (i.e., at the points P' ,

Q', R', \dots) and destructive interference occurs at $x = 0, \lambda/2, \lambda, 3\lambda/2, \dots$ (i.e., at points A, P, Q, R, \dots).** It may be mentioned that when constructive and destructive interferences occur, there is no violation of the principle of conservation of energy; the energy is merely redistributed.

In general, if we have n displacements

$$\left. \begin{aligned} x_1 &= a_1 \cos(\omega t + \theta_1) \\ x_2 &= a_2 \cos(\omega t + \theta_2) \\ \dots &\dots \dots \dots \dots \dots \dots \\ x_n &= a_n \cos(\omega t + \theta_n) \end{aligned} \right\} \quad (31)$$

then

$$x = x_1 + x_2 + \dots + x_n = a \cos(\omega t + \theta) \quad (32)$$

where

$$a \cos \theta = a_1 \cos \theta_1 + \dots + a_n \cos \theta_n \quad (33)$$

and

$$a \sin \theta = a_1 \sin \theta_1 + \dots + a_n \sin \theta_n \quad (34)$$

11.6 THE GRAPHICAL METHOD FOR STUDYING SUPERPOSITION OF SINUSOIDAL WAVES

In this section we will discuss the graphical method for adding displacements of the same frequency. This method is particularly useful when we have a large number of superposing waves as it indeed happens when we consider the phenomenon of diffraction.

Let us first try to obtain the resultant of the two displacements given by Eq. (20) using the graphical method. We draw a circle of radius a_1 and let the point P on the circle be such that OP makes an angle θ_1 with the x -axis*** (see Fig. 11.8). We next draw a circle of radius a_2 and let the point Q on the circle be such that OQ makes an angle θ_2 with the x -axis. We use the law of parallelograms to find the resultant \overrightarrow{OR} of the vectors \overrightarrow{OP} and \overrightarrow{OQ} . The length of the vector \overrightarrow{OR} will represent the amplitude of the resultant displacement and if θ is the angle that OR makes with the x -axis, then the initial phase of the resultant will be θ . This can be easily seen by noting that

$$\begin{aligned} OR \cos \theta &= OP \cos \theta_1 + PR \cos \theta_2 \\ &= a_1 \cos \theta_1 + a_2 \cos \theta_2 \end{aligned} \quad (35)$$

*Indeed in Sec. 11.2, while discussing stationary waves on a string, we had, at a particular value of x , two sinusoidal waves of the same frequency (but having different initial phases) superposing on each other. However, in general, one could have superposition of displacements which are in different directions; for example, the superposition of two linearly polarized waves to produce a circularly polarized wave (see Chapter 19).

**In Chapter 12, we will study the interference pattern produced by the superposition of spherical waves emanating from two point sources.

***Clearly, if we assume the vector \overrightarrow{OP} to rotate (in the anticlockwise direction) with angular velocity ω then the x -coordinate of the vector \overrightarrow{OP} will be $a_1 \cos(\omega t + \theta_1)$ where $t = 0$ corresponds to the instant when the rotating vector is at the point P .

Similarly,

$$OR \sin \theta = a_1 \sin \theta_1 + a_2 \sin \theta_2 \quad (36)$$

Thus, if we wish to find the resultant of the two displacements given by Eq. (20) then we must first draw a vector (\overrightarrow{OP}) of length a_1 making an angle θ_1 with the axis; from the tip of this vector we must draw another vector (\overrightarrow{PR}) of length a_2 making an angle θ_2 with the axis. The length of the vector \overrightarrow{OR} will represent the resultant amplitude and the angle that it makes with the axis will represent the initial phase of the resultant displacement. It can be easily seen that if we have a third displacement

$$x_3 = a_3 \cos (\omega t + \theta_3) \quad (37)$$

As an illustration of the above procedure we consider the resultant of N simple harmonic motions all having the same amplitude and with their phases increasing in arithmetic progression. Thus

[illegible]

In Fig. 11.9 the vectors $\overrightarrow{OP_1}$, $\overrightarrow{P_1P_2}$, $\overrightarrow{P_2P_3}$, ... correspond to x_1, x_2, x_3, \dots respectively. The resultant is denoted by the vector $\overrightarrow{OP_N}$. Let Q_1L and Q_2L be the perpendicular bisectors of OP_1 and P_1P_2 . It is easy to prove that

$$\Delta LQ_1P_1 \equiv \Delta LQ_2P_1$$

L and radius is LO . Further, $\angle LP_1O = \frac{\pi - \theta_0}{2}$ and, therefore, $\angle OLP_1 = \theta_0$. Thus,

$$LO = \frac{a/2}{\sin \theta_0/2}$$

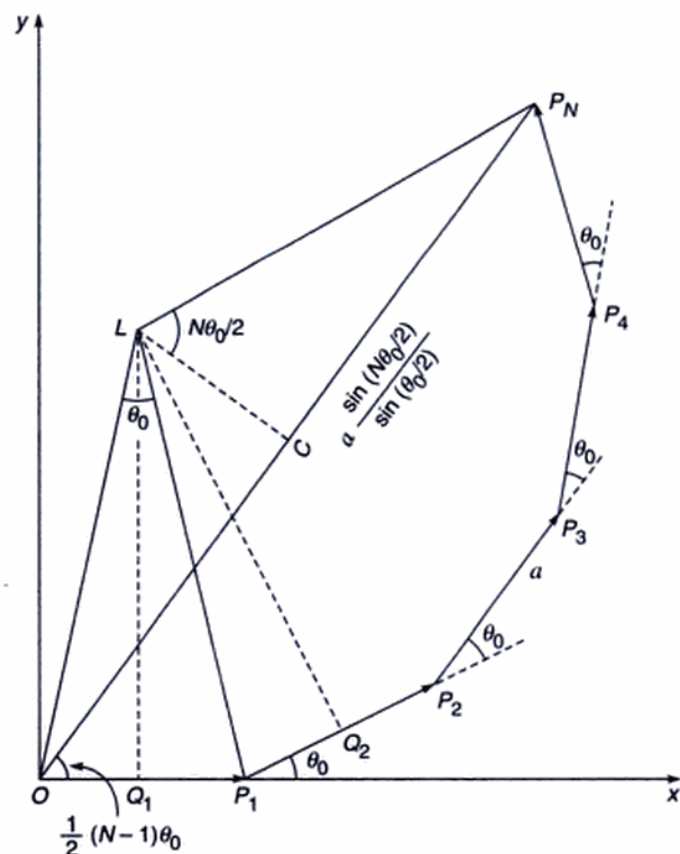


Fig. 11.9 The graphical method for determining the resultant of N simple harmonic motions along the same direction and having the same frequency.

and

$$OP_N = 2OC = 2LO \sin \frac{N\theta_0}{2}$$

$$= a \frac{\sin \frac{N\theta_0}{2}}{\sin \frac{\theta_0}{2}} \quad (39)$$

Further, the phase of the resultant displacement would be

$$\angle P_N OX = \frac{1}{2} (N-1)\theta_0$$

Thus

$$a \cos \omega t + a \cos (\omega t + \theta_0) + \dots + a \cos [\omega t + (N-1)\theta_0]$$

$$= A \cos (\omega t + \theta) \quad (40)$$

where

$$A = \frac{a \sin \frac{N\theta_0}{2}}{\sin \frac{\theta_0}{2}} \quad (41)$$

and

$$\theta = \frac{1}{2} (N-1)\theta_0 \quad (42)$$

We will use this result in Chapter 16.

11.7 THE COMPLEX REPRESENTATION

Often it is more convenient to use the complex representation in which the displacement

$$x_1 = a_1 \cos (\omega t + \theta_1) \quad (43)$$

is written as

$$x_1 = a_1 e^{i(\omega t + \theta_1)} \quad (44)$$

where it is implied that the actual displacement is the real part of x_1 . Further, if

$$x_2 = a_2 e^{i(\omega t + \theta_2)}$$

then

$$x_1 + x_2 = (a_1 e^{i\theta_1} + a_2 e^{i\theta_2}) e^{i\omega t} = a e^{i(\omega t + \theta)} \quad (45)$$

where

$$a e^{i\theta} = a_1 e^{i\theta_1} + a_2 e^{i\theta_2} \quad (46)$$

If we equate the real and imaginary parts of Eq. (46), we would obtain Eqs (23) and (24).

An interesting illustration of the usefulness of this method is to consider the resultant of the N displacements described by Eq. (38). Thus we write

$$x_1 = a e^{i\omega t}, x_2 = a e^{i(\omega t + \theta_0)}, \dots$$

Hence

$$x = x_1 + x_2 + \dots$$

$$= a e^{i\omega t} [1 + e^{i\theta_0} + e^{2i\theta_0} + \dots + e^{i(N-1)\theta_0}]$$

$$= a e^{i\omega t} \frac{1 - e^{Ni\theta_0}}{1 - e^{i\theta_0}}$$

$$= a e^{i\omega t} \frac{e^{iN\theta_0/2}}{e^{i\theta_0/2}} \cdot \frac{e^{iN\theta_0/2} - e^{-iN\theta_0/2}}{e^{i\theta_0/2} - e^{-i\theta_0/2}}$$

$$= \frac{a \sin \frac{N\theta_0}{2}}{\sin \frac{\theta_0}{2}} \exp \left[i \left\{ \omega t + (N-1) \frac{\theta_0}{2} \right\} \right] \quad (47)$$

which is consistent with Eq. (40). The complex representation is also very useful in considering the spreading of a wave packet (see Sec. 8.3).

It may be noted that whereas

$$\text{Re}(x_1) + \text{Re}(x_2) = \text{Re}(x_1 + x_2)$$

but

$$(\text{Re } x_1)(\text{Re } x_2) \neq \text{Re}(x_1 x_2)$$

where $\text{Re}(\dots)$ denotes the 'real part of' the quantity inside the brackets. Thus, one must be careful in calculating the intensity of a wave which is proportional to the square of the amplitude. While using the complex representation, one must calculate the amplitude first and then the intensity.

SUMMARY

- According to the principle of superposition of waves, the resultant displacement (at a particular point) produced by a number of waves is the vector sum of the displacements produced by each one of the disturbances.
- The stationary waves on a string and the formation of standing electromagnetic waves are formed by the superposition of waves traveling in opposite direction.
- If the two displacements (produced by two sinusoidal waves) are in phase, then the resultant amplitude will be the sum of the two amplitudes; this is known as constructive interference. On the other hand, if the two displacements are π out of phase, then the resultant amplitude will be the difference of the two amplitudes; this is known as destructive interference.

PROBLEMS

- 11.1 Standing waves are formed on a stretched string under tension of 1 Newton. The length of the string is 30 cm and it vibrates in 3 loops. If the mass per

unit length of the wire is 10 mg/cm, calculate the frequency of the vibrations.

- 11.2** In Problem 11.1, if the string is made to vibrate in its fundamental mode, what will be the frequency of vibration?

- 11.3** In the experimental arrangement of Wiener, what should be the angle between the film and the mirror if the distance between two consecutive dark bands is 7×10^{-3} cm. Assume $\lambda = 6 \times 10^{-5}$ cm.

[Ans: $\sim 1/4^\circ$]

- 11.4** Standing waves with five loops are produced on a stretched string under tension. The length of the string is 50 cm and the frequency of vibrations is 250 sec^{-1} . Calculate the time variation of the displacement of the points which are at distances of 2 cm, 5 cm, 15 cm, 18 cm, 20 cm, 35 cm and 45 cm from one end of the string.

- 11.5** The displacements associated with two waves (propagating in the same direction) having same amplitude but slightly different frequencies can be written in the form

$$a \cos 2\pi \left(vt - \frac{x}{\lambda} \right)$$

and
$$a \cos 2\pi \left((v + \Delta v)t - \frac{x}{(\lambda - \Delta \lambda)} \right)$$

(Such displacements are indeed obtained when we have two tuning forks with slightly different

frequencies.) Discuss the superposition of the displacements and show that at a particular value of x , the intensity will vary with time.

- 11.6** In Problem 11.5 assume $v = 330 \text{ m/sec}$, $\nu = 256 \text{ sec}^{-1}$, $\Delta \nu = 2 \text{ sec}^{-1}$ and $a = 0.1 \text{ cm}$. Plot the time variation of the intensity at $x = 0$, $\frac{\lambda}{4}$ and $\frac{\lambda}{2}$.

- 11.7** Use the complex representation to study the time variation of the resultant displacement at $x = 0$ in Problems 11.5 and 11.6.

- 11.8** Discuss the superposition of two plane waves (of the same frequency and propagating in the same direction) as a function of the phase difference between them. (Such a situation indeed arises when a plane wave gets reflected at the upper and lower surfaces of a glass slab; see Sec. 13.2.)

- 11.9** In Example 9.1 we had discussed the propagation of a semicircular pulse on a string. Consider two semicircular pulses propagating in opposite directions. At $t = 0$, the displacement associated with the pulses propagating in the $+x$ and in the $-x$ directions are given by

$$[R^2 - x^2]^{1/2} \text{ and } -[R^2 - (x - 10R)^2]^{1/2}$$

respectively. Plot the resultant disturbance at $t = R/v, 2.5R/v, 5R/v, 7.5R/v$ and $10R/v$; where v denotes the speed of propagation of the wave.

REFERENCES AND SUGGESTED READING

See at the end of Chapter 12.

Chapter 12

Two Beam Interference by Division of Wavefront

'The wave nature of light was demonstrated convincingly for the first time in 1801 by Thomas Young by a wonderfully simple experiment...He let a ray of sunlight into a dark room, placed a dark screen in front of it, pierced with two small pinholes, and beyond this, at some distance a white screen. He then saw two darkish lines at both sides of a bright line, which gave him sufficient encouragement to repeat the experiment, this time with spirit flame as light source, with a little salt in it, to produce the bright yellow sodium light. This time he saw a number of dark lines, regularly spaced; the first clear proof that light added to light can produce darkness. This phenomenon is called interference. Thomas Young had expected it because he believed in the wave theory of light.'

— Dennis Gabor in his Nobel Lecture, December 11, 1971

Thomas Young had amazing broad interests and talents ... From his discoveries in medicine and science, Helmholtz concluded: 'His was one of the most profound minds that the world has ever seen.'

— From the Internet

12.1 INTRODUCTION

In the previous chapter, we had considered the superposition of one-dimensional waves propagating on a string and had shown that there is a variation of energy density along the length of the string due to the interference of two waves (see Fig. 11.5). In general, whenever two waves superpose, one obtains an intensity distribution which is known as the interference pattern. In this chapter, we will consider the interference pattern produced by waves emanating from two point sources. It may be mentioned that with sound waves the interference pattern can be observed without much difficulty because the two interfering waves maintain a constant phase relationship; this is also the case for microwaves. However, for light waves, due to the very process of emission, one cannot observe interference between the waves from two independent sources,* although the interference does take place (see Sec. 12.4). Thus, one tries to derive interfering waves from a single wave so that the phase relationship is maintained. The methods to achieve this can be classified under two broad categories. Under the first category, in a typical arrangement, a beam is allowed to fall on two closely spaced holes and the two beams emanating from the holes interfere. This method is known as division of wavefront and will be discussed in detail in this chapter. In the other method, known as division of amplitude, a beam is divided at two or more reflecting surfaces

and the reflected beams interfere. This will be discussed in the next chapter. We must, however, emphasize that the present and the following chapters are based on one underlying principle, namely the superposition principle.

It may be mentioned that it is also possible to observe interference using multiple beams; this is known as multiple beam interferometry and will be discussed in Chapter 14. It will be shown that multiple beam interferometry offers some unique advantages over two beam interferometry.

12.2 INTERFERENCE PATTERN PRODUCED ON THE SURFACE OF WATER

We consider surface waves emanating from two point sources in a water tank. We may have, for example, two sharp needles vibrating up and down at the points S_1 and S_2 (see Fig. 12.1). Although water waves are not really transverse, we will, for the sake of simplicity, assume water waves to produce displacements which are transverse to the direction of propagation.

If there was only one needle (say at S_1) vibrating with a certain frequency ν then circular ripples would have spread out from the point S_1 . The wavelength would have been v/ν and the crests and troughs would have moved outwards. Similarly for the vibrating needle at S_2 . However, if both

*It is difficult to observe the interference pattern even with two laser beams unless they are phase locked.

needles are vibrating, then waves emanating from S_1 will interfere with the waves emanating from S_2 . We assume that the needle at S_2 vibrates in phase with the needle at S_1 ; i.e., S_1 and S_2 go up simultaneously, they also reach the lowest position at the same time. Thus, if at a certain instant, the disturbance emanating from the source S_1 produced a crest at a distance ρ from S_1 then the disturbance from S_2 would also produce a crest at a distance ρ from S_2 , etc. This is explicitly shown in Fig. 12.1, where the solid curves represent (at a particular instant) the positions of the crests due to disturbances emanating from S_1 and S_2 . Similarly, the dashed curves represent (at the same instant) the positions of the troughs. Notice that at all points on the perpendicular bisector OY the disturbances reaching from S_1 and from S_2 will always be in phase. Consequently, at an arbitrary point A (on the perpendicular bisector) we may write the resultant disturbance as

$$\begin{aligned} y &= y_1 + y_2 \\ &= 2a \cos \omega t \end{aligned} \quad (1)$$

where $y_1 (= a \cos \omega t)$ and $y_2 (= a \cos \omega t)$ represent the displacements at the point A due to S_1 and S_2 respectively. We see that the amplitude at A is twice the amplitude produced by each one of the source. It should be noted that at $t = \frac{T}{4} (= \frac{1}{4v} = \frac{\pi}{2\omega})$ the displacements produced at the point A by each of the source would be zero and the resultant will also be zero. This is also obvious from Eq. (1).

Next, let us consider a point B such that

$$S_2B - S_1B = \lambda/2 \quad (2)$$

At such a point the disturbance reaching from the source S_1 will always be out of phase with the disturbance reaching from S_2 . This follows from the fact that the disturbance reaching the point B from the source S_2 must have started

half a period ($= T/2$) earlier than the disturbance reaching B from S_1 . Consequently, if the displacement at B due to S_1 is given by

$$y_1 = a \cos \omega t$$

then the displacement at B due to S_2 would be given by

$$y_2 = a \cos (\omega t - \pi) = -a \cos \omega t$$

and the resultant $y = y_1 + y_2$ is zero at all times. Such a point corresponds to destructive interference and is known as a node and corresponds to minimum intensity. It may be pointed out that the amplitudes of the two vibrations reaching the point B will not really be equal as it is at different distances from S_1 and S_2 . However, if the distances involved are large (in comparison to the wavelength), the two amplitudes will be very nearly equal and the resultant intensity will be very nearly zero.

In a similar manner we may consider a point C such that

$$S_2C - S_1C = \lambda$$

where the phase of the vibrations (reaching from S_1 and S_2) are exactly the same as at the point A . Consequently we will again have constructive interference. In general, if a point P is such that

$$S_2P - S_1P = n\lambda \text{ (maxima)} \quad (3)$$

$n = 0, 1, 2, \dots$, then the disturbances reaching the point P from the two sources will be in phase, the interference will be constructive and the intensity will be maximum. On the other hand, if the point P is such that

$$S_2P - S_1P = \left(n + \frac{1}{2}\right)\lambda \text{ (minima)} \quad (4)$$

then the disturbances reaching the point P from the two sources will be out of phase, the interference will be destructive and the intensity will be minimum. The actual

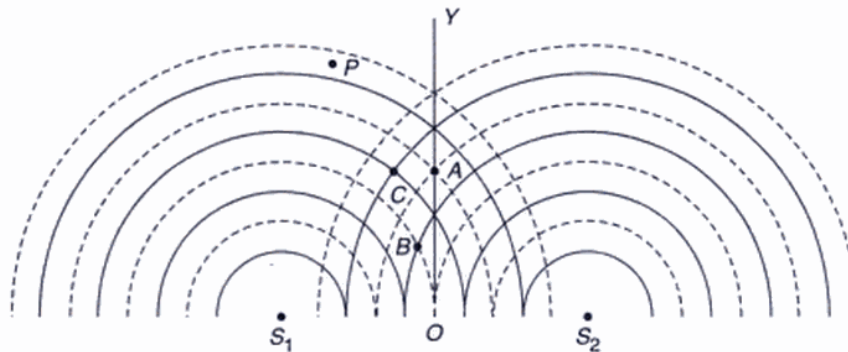


Fig. 12.1 Waves emanating from two point sources S_1 and S_2 vibrating in phase. The solid and the dashed curves represent the positions of the crests and troughs respectively.

interference pattern produced from two point sources vibrating in phase in a ripple tank is shown in Fig. 12.2.

Example 12.1 The intensity at the point which neither satisfies Eq. (3) nor Eq. (4) will neither be a maximum nor zero. Consider a point P such that $S_2P - S_1P = \lambda/3$. Find the ratio of the intensity at the point P to that at a maximum.

Solution: If the disturbance reaching the point P from S_1 is given by

$$y_1 = a \cos \omega t$$

then the disturbance from S_2 would be given by

$$y_2 = a \cos \left(\omega t - \frac{2\pi}{3} \right)$$

because a path difference of $\lambda/3$ corresponds to a phase difference of $\frac{2\pi}{3}$.

Thus the resultant displacement would be

$$\begin{aligned} y &= y_1 + y_2 \\ &= a \left[\cos \omega t + \cos \left(\omega t - \frac{2\pi}{3} \right) \right] \\ &= 2a \cos \left(\omega t - \frac{\pi}{3} \right) \cos \frac{\pi}{3} \\ &= a \cos \left(\omega t - \frac{\pi}{3} \right) \end{aligned}$$

The intensity is therefore 1/4th of the intensity at the maxima. In a similar manner one can calculate the intensity at any other point.

Example 12.2 The locus of points which correspond to minima are known as nodal lines. Show that the equation of a nodal line is a hyperbola. Also obtain the locus of points which correspond to maxima.

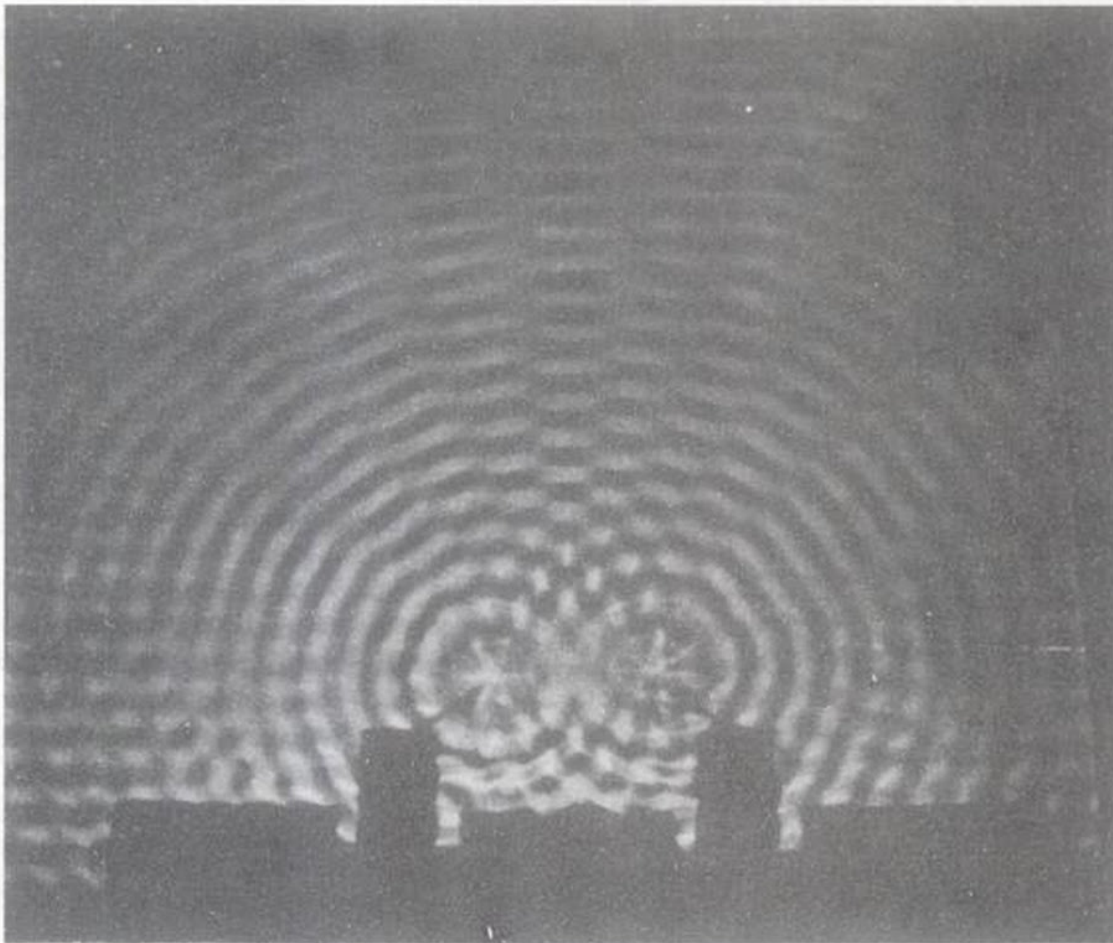


Fig. 12.2 The actual interference pattern produced from two point sources vibrating in phase in a ripple tank (After Ref. 9, used with permission).

Solution: For the sake of generality we find the locus of the point P which satisfies the following equation:

$$S_1P - S_2P = \Delta \quad (5)$$

Thus, if $\Delta = n\lambda$, we have a maximum and if $\Delta = (n + \frac{1}{2})\lambda$ we have a minimum. We choose the midpoint of S_1S_2 as the origin, with the x -axis along S_1S_2 and the y -axis perpendicular to it (see Fig. 12.3). If the distance between S_1 and S_2 is d , then the coordinates of the points S_1 and S_2 are $(-\frac{d}{2}, 0)$ and $(+\frac{d}{2}, 0)$ respectively. Let the coordinates of the point P be (x, y) . Then

$$S_1P = \left[\left(x + \frac{d}{2} \right)^2 + y^2 \right]^{1/2}$$

and

$$S_2P = \left[\left(x - \frac{d}{2} \right)^2 + y^2 \right]^{1/2}$$

Therefore,

$$S_1P - S_2P = \left[\left(x + \frac{d}{2} \right)^2 + y^2 \right]^{1/2} - \left[\left(x - \frac{d}{2} \right)^2 + y^2 \right]^{1/2} = \Delta$$

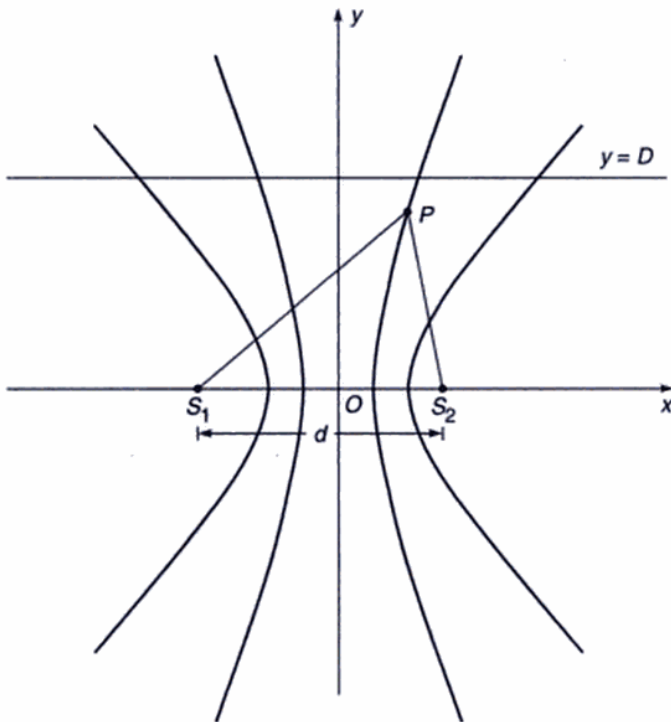


Fig. 12.3 The nodal curves.

or

$$\left(x + \frac{d}{2} \right)^2 + y^2 = \Delta^2 + \left(x - \frac{d}{2} \right)^2 + y^2 + 2\Delta \left[\left(x - \frac{d}{2} \right)^2 + y^2 \right]^{1/2}$$

or

$$2xd - \Delta^2 = 2\Delta \left[\left(x - \frac{d}{2} \right)^2 + y^2 \right]^{1/2}$$

On squaring, we obtain

$$4x^2d^2 - 4xd\Delta^2 + \Delta^4 = 4\Delta^2 \left[x^2 - xd + \frac{d^2}{4} + y^2 \right]$$

Thus we obtain

$$\frac{x^2}{\frac{1}{4}\Delta^2} - \frac{y^2}{\frac{1}{4}(d^2 - \Delta^2)} = 1 \quad (6)$$

which is the equation of a hyperbola. When $\Delta = (n + \frac{1}{2})\lambda$ the curves correspond to minima and when $\Delta = n\lambda$ the curves correspond to maxima. For large values of x and y the curves asymptotically tend to the straight lines

$$y = \pm \left(\frac{d^2 - \Delta^2}{\Delta^2} \right)^{1/2} x \quad (7)$$

It may be pointed out that there is no point P for which $S_1P \sim S_2P > d$ ($S_1P \sim S_2P$ equals d on the x -axis only). Now, it appears from Eq. (6) that when $\Delta > d$, the resulting equation is an ellipse which we know is impossible. The fallacy is a result of the fact that because of a few squaring operations, Eq. (6) also represents the locus of all those points for which $S_1P + S_2P = \Delta$ and obviously in this case Δ can exceed d .

Example 12.3 Consider a line parallel to the x -axis at a distance D from the origin (see Fig. 12.3). Assume $D \gg \lambda$. Find the points on this line where minimum intensity will occur.

Solution: The equation of this line would be

$$y = D \quad (8)$$

Further at large distances from the origin the equation of the nodal lines would be

$$y = \pm \left(\frac{d^2 - \Delta_n^2}{\Delta_n^2} \right)^{1/2} x \quad (9)$$

where $\Delta_n = (n + \frac{1}{2})\lambda$; $n = 0, 1, 2, \dots$. Clearly the points at

which minima will occur (on the line $y = D$) would be given by

$$\begin{aligned} x_n &= \pm \left(\frac{\Delta_n^2}{d^2 - \Delta_n^2} \right)^{1/2} D \\ &= \pm \frac{\Delta_n}{d} \left(1 - \frac{\Delta_n^2}{d^2} \right)^{-1/2} D \\ &\approx \pm \left(n + \frac{1}{2} \right) \frac{\lambda D}{d} \end{aligned} \quad (10)$$

where we have assumed $\Delta_n \ll d$. Thus the points corresponding to minima will be equally spaced with a spacing of $\lambda D/d$.

Example 12.4 Till now we have assumed the needles at S_1 and S_2 (see Fig. 12.1) to vibrate in phase. Assume now that the needles vibrate with a phase difference of π and obtain the nodal lines. Generalize the result for an arbitrary phase difference between the vibrations of the two needles.

Solution: The two needles S_1 and S_2 vibrate out of phase. Thus if, at any instant, the needle at S_1 produces a crest at a distance R from it then the needle at S_2 would produce a trough at a distance R from S_2 . Therefore, at all points on the perpendicular bisector OY (see Fig. 12.4) the two vibrations will always be out of phase and we will have minimum. On the other hand, at the point B which satisfies the equation

$$S_2B - S_1B = \lambda/2$$

the two vibrations will be in phase and we will have maximum. Thus, because of the initial phase difference of π , the conditions for maxima and minima are reversed, i.e., when

$$S_2P \sim S_1P = \left(n + \frac{1}{2} \right) \lambda \text{ (maxima)}$$

the interference will be constructive and we will have maxima, and when

$$S_2P \sim S_1P = n\lambda \text{ (minima)}$$

the interference will be destructive and we will have minima. Notice that one again obtains a stationary interference pattern with nodal lines as hyperbolae.

The above analysis can easily be generalized for arbitrary phase difference between the two needles. Assume, for example, that there is a phase difference of $\pi/3$, i.e., if there is a crest at a distance R from S_1 then there is a crest at a distance $R - \lambda/6$ from S_2 . Consequently, the condition

$$S_1P - S_2P = n\lambda + \frac{\lambda}{6}; n = 0, \pm 1, \pm 2, \dots$$

will correspond to maxima.

12.3 COHERENCE

From the above examples we find that whenever the two needles vibrate with a constant phase difference, a stationary interference pattern is produced. The positions of the maxima and minima will, however, depend on the phase difference in the vibration of the two needles. Two sources which vibrate with a fixed phase difference between them are said to be *coherent*.

We next assume that the two needles are sometimes vibrating in phase, sometimes vibrating out of phase, sometimes vibrating with a phase difference of $\pi/3$, etc., then the interference pattern will keep on changing. If the phase difference changes with such great rapidity that a stationary interference cannot be observed then the sources are said to be *incoherent*.

Let the displacement produced by the sources at S_1 and S_2 be given by

$$\left. \begin{aligned} y_1 &= a \cos \omega t \\ y_2 &= a \cos (\omega t + \phi) \end{aligned} \right\} \quad (11)$$

then the resultant displacement would be

$$y = y_1 + y_2 = 2a \cos \phi/2 \cos (\omega t + \phi/2) \quad (12)$$

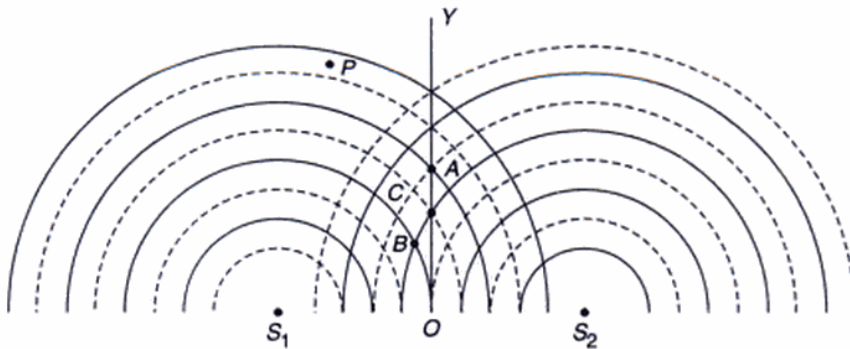


Fig. 12.4 Waves emanating from two point sources S_1 and S_2 vibrating out of phase.

The intensity (I) which is proportional to the square of the amplitude can be written in the form

$$I = 4I_0 \cos^2 \phi/2 \quad (13)$$

where I_0 is the intensity produced by each one of the source individually. Clearly if $\phi = \pm\pi, \pm3\pi, \dots$, the resultant intensity will be zero and we will have minima. On the other hand, when $\phi = 0, \pm2\pi, \pm4\pi, \dots$, the intensity will be maximum ($= 4I_0$). However, if the phase difference between the sources S_1 and S_2 (i.e., ϕ) is changing with time, the observed intensity will be given by

$$I = 4I_0 \left\langle \cos^2 \frac{\phi}{2} \right\rangle \quad (14)$$

where $\langle \dots \rangle$ denotes the time average of the quantity inside the angular brackets; the time average of a time dependent function is defined by the following relation:

$$\langle f(t) \rangle = \frac{1}{\tau} \int_{-\tau/2}^{+\tau/2} f(t) dt \quad (15)$$

where τ represents the time over which the averaging is carried out. For example, if the interference pattern is viewed by a normal eye, this averaging will be over about 1/10th of a second; for a camera with exposure time 0.001 sec, $\tau = 0.001$ sec. etc. Clearly, if ϕ varies in a random manner in times which are small compared to τ , then $\cos^2 \phi/2$ will randomly vary between 0 and 1 and $\langle \cos^2 \phi/2 \rangle$ would be 1/2 [see also Sec. 12.6]. For such a case

$$I = 2I_0 \quad (16)$$

which implies that if the sources are incoherent then the resultant intensity is the sum of the two intensities and there is no variation of intensity! Thus, if one (or both) of the two vibrating sources are turned on and off in a random manner (such that the phase difference between the vibrations of the two sources varies rapidly) then the interference phenomenon will not be observed. We will discuss this point again in Sec. 12.6 and also in Chapter 15.

12.4 INTERFERENCE OF LIGHT WAVES

Till now we have considered interference of waves produced on the surface of water. We will now discuss the

interference pattern produced by light waves; however, for light waves it is difficult to observe a stationary interference pattern. For example, if we use two conventional light sources (like two sodium lamps) illuminating two pinholes (see Fig. 12.5), we will not observe any interference pattern on the screen. This can be understood from the following reasoning: In a conventional light source, light comes from a large number of independent atoms; each atom emitting light for about 10^{-10} sec, i.e., light emitted by an atom is essentially a pulse lasting for only 10^{-10} sec.* Even if the atoms were emitting under similar conditions, waves from different atoms would differ in their initial phases.

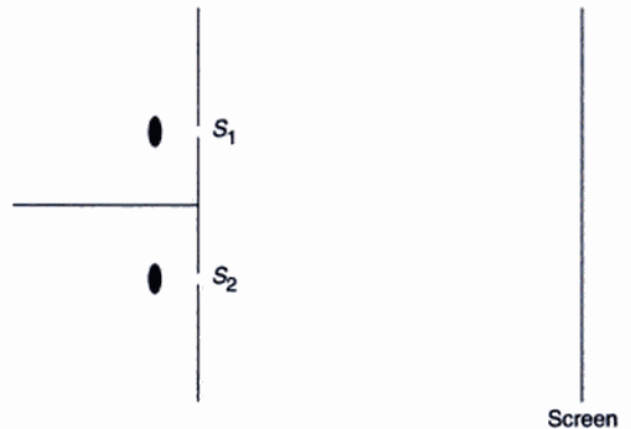


Fig. 12.5 If two sodium lamps illuminate two pinholes S_1 and S_2 , no interference pattern will be observed on the screen.

Consequently, light coming out from the holes S_1 and S_2 will have a fixed phase relationship for a period of about 10^{-10} sec, hence the interference pattern will keep on changing every billionth of a second. The eye can notice intensity changes which last at least for a tenth of a second and hence we will observe a uniform intensity over the screen. However, if we have a camera whose time of shutter opening can be made less than 10^{-10} sec then the film will record an interference pattern**. We summarise the above results by noting that light beams from two independent sources do not have any fixed relationship***, as such they do not produce any stationary interference pattern.

Thomas Young in 1801 devised an ingenious but simple method to lock the phase relationship between the two sources. The trick lies in the division of a single wavefront into two; these two split wavefronts act as if they emanated

*Since the optical frequencies are of the order of 10^{15} sec^{-1} , such a short pulse consists of about a million oscillations; thus it is almost monochromatic (see Chapter 15).

**This interference pattern will be a set of dark and bright bands only if the light waves have the same state of polarization. This can, however, be easily done by putting two polaroids in front of S_1 and S_2 . We should mention here that by using two independent laser beams it has been possible to record the interference pattern (see Chapter 15).

***Such sources are termed as incoherent sources.

from two sources having a fixed phase relationship and, therefore, when these two waves were allowed to interfere, a stationary interference pattern was obtained. In the actual experiment a light source illuminates the pinhole S (see Fig. 12.6). Light diverging from this pinhole fell on a barrier which contained two pinholes S_1 and S_2 which were very close to one another and were located equidistant from S . Spherical waves emanating from S_1 and S_2 (see Fig. 12.7) were coherent and on the screen beautiful interference fringes were obtained. In order to show that this was indeed an interference effect, Young showed that the fringes on the screen disappear when S_1 (or S_2) is covered up. Young explained the interference pattern by considering the principle of superposition, and by measuring the distance between the fringes he calculated the wavelength. Figure 12.7 shows the section of the wavefront on the plane containing S , S_1 and S_2 .

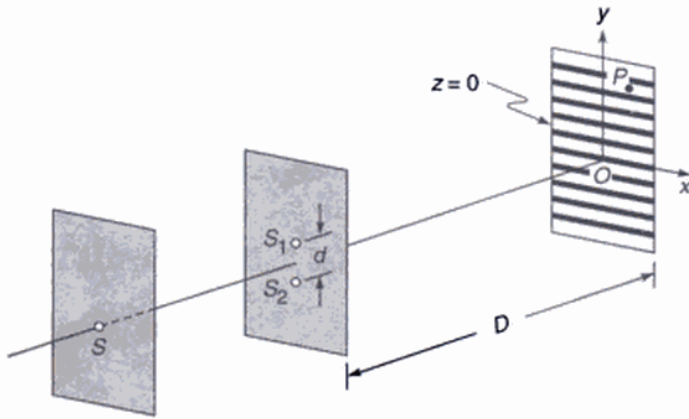


Fig. 12.6 Young's arrangement to produce interference pattern.

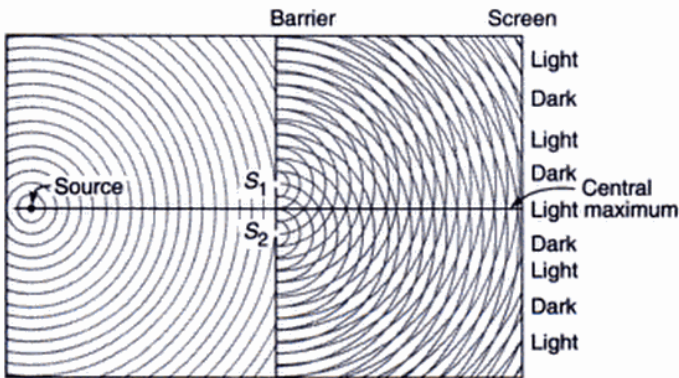


Fig. 12.7 Sections of the spherical wavefronts emanating from S , S_1 and S_2 (Adapted from Ref. 7; used with permission).

12.5 THE INTERFERENCE PATTERN

Let S_1 and S_2 represent the two pinholes of the Young's interference experiment. We would determine the positions

of maxima and of minima on the line LL' which is parallel to the y -axis and lies in the plane containing the points S , S_1 and S_2 (see Fig. 12.8). We will show that the interference pattern (around the point O) consists of a series of dark and bright lines perpendicular to the plane of Fig. 12.8; O being the foot of the perpendicular from the point S on the screen.

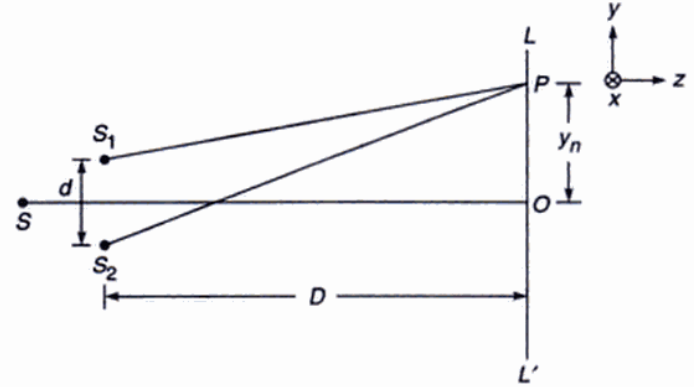


Fig. 12.8 Arrangement for producing Young's interference pattern.

For an arbitrary point P (on the line LL') to correspond to a maximum we must have

$$S_2P - S_1P = n\lambda; \quad n = 0, 1, 2, \dots \quad (17)$$

Now,

$$\begin{aligned} (S_2P)^2 - (S_1P)^2 &= \left[D^2 + \left(y_n + \frac{d}{2} \right)^2 \right] - \left[D^2 + \left(y_n - \frac{d}{2} \right)^2 \right] \\ &= 2y_nd \end{aligned}$$

where

$$S_1S_2 = d \quad \text{and} \quad OP = y_n$$

Thus

$$S_2P - S_1P = \frac{2y_nd}{S_2P + S_1P} \quad (18)$$

If $y_n, d \ll D$ then negligible error will be introduced if $S_2P + S_1P$ is replaced by $2D$. For example, for $d = 0.02$ cm, $D = 50$ cm, $OP = 0.5$ cm (which corresponds to typical values for a light interference experiment)

$$\begin{aligned} S_2P + S_1P &= [(50)^2 + (0.51)^2]^{1/2} + [(50)^2 + (0.49)^2]^{1/2} \\ &\approx 100.005 \text{ cm} \end{aligned}$$

Thus if we replace $S_2P + S_1P$ by $2D$, the error involved is about 0.005%. In this approximation, Eq. (18) becomes

$$S_2P - S_1P \approx \frac{y_nd}{D} \quad (19)$$

Using Eq. (17) we obtain

$$y_n = \frac{n\lambda D}{d} \quad (20)$$

Thus the dark and bright fringes are equally spaced and the distance between two consecutive dark (or bright) fringes is given by

$$\beta = y_{n+1} - y_n = \frac{(n+1)\lambda D}{d} - \frac{n\lambda D}{d}$$

or

$$\beta = \frac{\lambda D}{d} \quad (21)$$

which is the expression for the fringe width.

In order to determine the shape of the interference pattern we first note that the locus of the point P such that

$$S_2P - S_1P = \Delta \quad (22)$$

is a hyperbola in any plane containing the points S_1 and S_2 (see Example 12.2). Consequently, the locus is a hyperbola of revolution obtained by rotating the hyperbola about the axis S_1S_2 . In order to find the shape of the fringe on the screen we assume the origin to be at the point O and the z -axis to be perpendicular to the plane of the screen as shown in Fig. 12.6. The y -axis is assumed to be parallel to S_2S_1 . We consider an arbitrary point P on the plane of the screen (i.e., $z = 0$) (see Fig. 12.6) Let its coordinates be $(x, y, 0)$. The coordinates of the points S_1 and S_2 are $(0, \frac{d}{2}, D)$

and $(0, -\frac{d}{2}, D)$ respectively. Thus

$$\begin{aligned} S_2P - S_1P &= \left[x^2 + \left(y + \frac{d}{2} \right)^2 + D^2 \right]^{1/2} - \left[x^2 + \left(y - \frac{d}{2} \right)^2 + D^2 \right]^{1/2} \\ &= \Delta \text{ (say)} \end{aligned}$$

or

$$\begin{aligned} &\left[x^2 + \left(y + \frac{d}{2} \right)^2 + D^2 \right] \\ &= \left\{ \Delta + \left[x^2 + \left(y - \frac{d}{2} \right)^2 + D^2 \right]^{1/2} \right\}^2 \end{aligned}$$

$$\text{or} \quad [2yd - \Delta^2]^2 = (2\Delta)^2 \left[x^2 + \left(y - \frac{d}{2} \right)^2 + D^2 \right]$$

Hence,

$$(d^2 - \Delta^2)y^2 - \Delta^2x^2 = \Delta^2 \left[D^2 + \frac{1}{4}(d^2 - \Delta^2) \right]$$

which is the equation of a hyperbola. Thus the shape of the fringes is hyperbolic. On rearranging we get

$$y = \pm \left(\frac{\Delta^2}{d^2 - \Delta^2} \right)^{1/2} \left[x^2 + D^2 + \frac{1}{4}(d^2 - \Delta^2) \right]^{1/2} \quad (23)$$

For values of x such that

$$x^2 \ll D^2 \quad (24)$$

the loci are straight lines parallel to the x -axis. Thus we obtain approximately straight line fringes on the screen. It should be emphasized that the fringes are straight lines although the sources S_1 and S_2 are point sources. It is easy to see that if we had slits instead of the point sources we would have obtained again straight line fringes with increased intensities.

The fringes so produced are said to be non-localized; they can be photographed by just placing a film on the screen; they can also be seen through an eyepiece.

12.6 THE INTENSITY DISTRIBUTION

Let E_1 and E_2 be the electric fields produced at the point P by S_1 and S_2 respectively (see Fig. 12.8). The electric fields E_1 and E_2 will, in general, have different directions and different magnitudes. However, if the distances S_1P and S_2P are very large in comparison to the distance S_1S_2 , the two fields will almost be in the same direction. Thus, we may write

$$\begin{aligned} E_1 &= \hat{i} E_{01} \cos \left(\frac{2\pi}{\lambda} S_1P - \omega t \right) \\ \text{and} \quad E_2 &= \hat{i} E_{02} \cos \left(\frac{2\pi}{\lambda} S_2P - \omega t \right) \end{aligned} \quad (25)$$

where \hat{i} represents the unit vector along the direction of either of the electric fields. The resultant field will be given by

$$\begin{aligned} E &= E_1 + E_2 \\ &= \hat{i} \left[E_{01} \cos \left(\frac{2\pi}{\lambda} S_1P - \omega t \right) + E_{02} \cos \left(\frac{2\pi}{\lambda} S_2P - \omega t \right) \right] \end{aligned} \quad (26)$$

The intensity (I) will be proportional to the square of the electric field and will be given by

$$I = KE^2 \quad (27)$$

or

$$I = K \left[E_{01}^2 \cos^2 \left(\frac{2\pi}{\lambda} S_1 P - \omega t \right) + E_{02}^2 \cos^2 \left(\frac{2\pi}{\lambda} S_2 P - \omega t \right) + E_{01} E_{02} \left\{ \cos \left[\frac{2\pi}{\lambda} (S_2 P - S_1 P) \right] + \cos \left[2\omega t - \frac{2\pi}{\lambda} (S_2 P + S_1 P) \right] \right\} \right] \quad (28)$$

where K is a proportionality constant.* For an optical beam the frequency is very large ($\omega \approx 10^{15} \text{ sec}^{-1}$) and all the terms depending on ωt will vary with extreme rapidity (10^{15} times in a second); consequently, any detector would record an average value of various quantities. Now

$$\begin{aligned} \langle \cos^2(\omega t - \theta) \rangle &= \frac{1}{2\tau} \int_{-\tau}^{+\tau} \frac{1 + \cos[2(\omega t - \theta)]}{2} dt \\ &= \frac{1}{2} + \frac{1}{16\pi} \frac{T}{\tau} \left[\sin 2(\omega t - \theta) \right]_{-\tau}^{+\tau} \end{aligned}$$

where $T = \frac{2\pi}{\omega} (\approx 2\pi \times 10^{-15} \text{ sec for an optical beam})$. For any practical detector** $\frac{T}{\tau} \ll 1$ and since the quantity between the curly brackets will always be between -2 and $+2$, we may write

$$\langle \cos^2(\omega t - \theta) \rangle \approx \frac{1}{2} \quad (29)$$

The factor $\cos(2\omega t - \phi)$ will oscillate between $+1$ and -1 and its average will be zero as can indeed be shown mathematically. Thus the intensity, that a detector will record, will be given by

$$I = I_1 + I_2 + 2\sqrt{I_1 I_2} \cos \delta \quad (30)$$

where

$$\delta = \frac{2\pi}{\lambda} (S_2 P - S_1 P) \quad (31)$$

represents the phase difference between the displacements reaching the point P from S_1 and S_2 . Further

$$I_1 = \frac{1}{2} K E_{01}^2$$

represents the intensity produced by the source S_1 if no light from S_2 is allowed to fall on the screen; similarly $I_2 = \frac{1}{2} K E_{02}^2$ represents the intensity produced by the source S_2 if no light from S_1 is allowed to fall on the screen. From Eq. (30) we may deduce the following:

- (a) The maximum and minimum values of $\cos \delta$ are $+1$ and -1 respectively; as such the maximum and minimum values of I are given by

$$\begin{aligned} I_{\max} &= (\sqrt{I_1} + \sqrt{I_2})^2 \\ I_{\min} &= (\sqrt{I_1} - \sqrt{I_2})^2 \end{aligned} \quad (32)$$

The maximum intensity occurs when

$$\delta = 2n\pi, \quad n = 0, 1, 2, \dots$$

or

$$S_2 P - S_1 P = n\lambda,$$

and the minimum intensity occurs when

$$\delta = (2n + 1)\pi; \quad n = 0, 1, 2, \dots$$

or

$$S_2 P - S_1 P = \left(n + \frac{1}{2} \right) \lambda$$

Notice that when $I_1 = I_2$, the intensity minimum is zero. In general, $I_1 \neq I_2$ and the minimum intensity is not zero.

- (b) If the holes S_1 and S_2 are illuminated by different light sources (see Fig. 12.4), then the phase difference δ will remain constant for about 10^{-10} sec (see discussion in Sec. 12.3) and thus δ would also vary with time*** in a random way. If we now carry out the averaging over time scales which are of the order of 10^{-8} sec , then

$$\langle \cos \delta \rangle = 0$$

and we obtain

$$I = I_1 + I_2$$

Thus, for two incoherent sources, the resultant intensity is the sum of the intensities produced by each one of the sources independently and no interference pattern is observed.

- (c) In the arrangement shown in Fig. 12.6, if the distances $S_1 P$ and $S_2 P$ are extremely large in comparison to d , then

$$I_1 = I_2 = I_0 \quad (\text{say})$$

*Eq. (27) will be derived in Chapter 19. In free space the constant K will be shown to be equal to $\epsilon_0 c^2$ where $\epsilon_0 (=8.854 \times 10^{-12} \text{ coul}^2/\text{N-m}^2)$ represents the permittivity of free space and c , the speed of light in free space.

**For a normal eye, $\tau \approx 0.1 \text{ sec}$; thus $T/\tau \approx 6 \times 10^{-14}$; even for a detector having 1 nsec as the resolution time, $T/\tau \approx 6 \times 10^{-5}$.

***Notice that this variation occurs in times of the order of 10^{-10} sec which is about a million times longer than the times for variation of the intensity due to the terms depending on ωt . Thus we are justified in first carrying out the averaging which leads to Eq. (30).

and

$$I = 2I_0 + 2I_0 \cos \delta = 4I_0 \cos^2 \frac{\delta}{2} \quad (33)$$

The intensity distribution (which is often termed as the \cos^2 pattern) is shown in Fig. 12.9. The actual fringe pattern (as it will appear on the screen) is

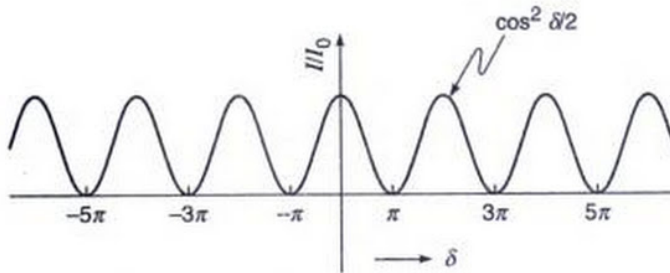


Fig. 12.9 The variation of intensity with δ .

shown in Fig. 12.10. Figures 12.10 (a) and (b) correspond to $d = 0.005$ mm ($\beta = 5$ mm) and $d = 0.025$ mm ($\beta = 1$ mm) respectively. Both figures correspond to $D = 5$ cm and $\lambda = 5 \times 10^{-5}$ cm. The values of the parameters are such that one can see the hyperbolic nature of the fringe pattern in Fig. 12.10(a).

Example 12.5 Instead of considering two point sources, we consider the superposition of two plane waves as shown in Fig. 12.11(a). The wave vectors for the two waves are given by

$$\mathbf{k}_1 = -\hat{y} k \sin \theta_1 + \hat{z} k \cos \theta_1$$

and

$$\mathbf{k}_2 = +\hat{y} k \sin \theta_2 + \hat{z} k \cos \theta_2$$

where $k = 2\pi/\lambda$ and θ_1 and θ_2 are defined in Fig. 12.11(a). Thus the electric fields of the two waves are described by the following two equations

$$\begin{aligned} E_1 &= E_{01} \cos [\mathbf{k}_1 \cdot \mathbf{r} - \omega t] \\ &= E_{01} \cos [-k y \sin \theta_1 + k z \cos \theta_1 - \omega t] \\ E_2 &= E_{02} \cos [\mathbf{k}_2 \cdot \mathbf{r} - \omega t] \\ &= E_{02} \cos [k y \sin \theta_2 + k z \cos \theta_2 - \omega t] \end{aligned}$$

If we assume $E_{01} = E_{02} = E_0$ and $\theta_1 = \theta_2 = \theta$ then the resultant field will be given by

$$E = 2 E_0 \cos (k y \sin \theta) \cos (k z \cos \theta - \omega t)$$

Thus the intensity distribution on the photograph plate LL' will be given by

$$I = 4 I_0 \cos^2(k y \sin \theta)$$

and the fringe pattern will be strictly straight lines with fringe width given by

$$\beta = \frac{\lambda}{2 \sin \theta}$$

Figure 12.11(b) shows the computer generated interference pattern on the screen LL' for $\theta = \pi/6$ and $\lambda = 5000 \text{ \AA}$. Thus $\beta = \lambda = 0.0005$ mm.

Example 12.6 In this example, we once again consider the interference pattern produced by 2 point sources S_1 and

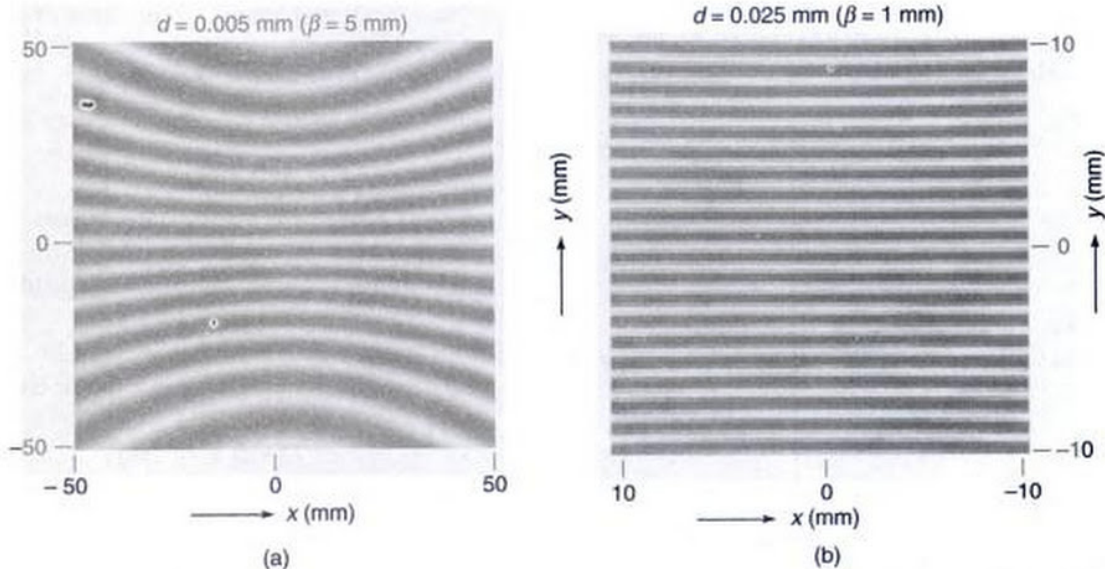


Fig. 12.10 Computer generated fringe pattern produced by two point sources S_1 and S_2 on the screen LL' (see Fig. 12.8); (a) and (b) correspond to $d = 0.005$ mm and 0.025 mm respectively (both figures correspond to $D = 5$ cm and $\lambda = 5 \times 10^{-5}$ cm).

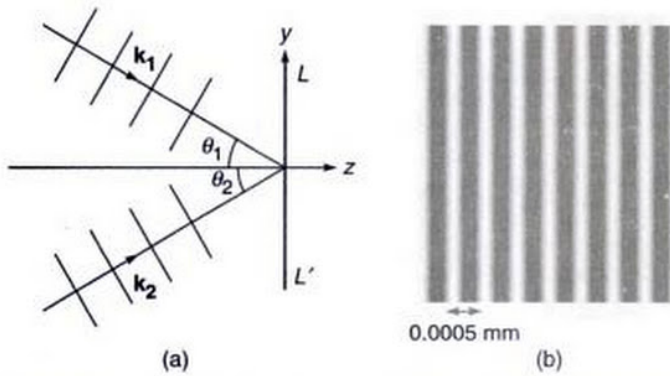


Fig. 12.11 (a) The superposition of two plane waves on LL' . (b) Computer generated interference pattern on the screen LL' for $\theta_1 = \theta_2 = \pi/6$ and $\lambda = 5000 \text{ \AA}$. The fringes are parallel to the x -axis.

S_2 on a plane PP' which is perpendicular to the line joining S_1 and S_2 [see Fig. 12.12(a)]. Obviously, on the plane PP' , the locus of the point P for which

$$S_1P - S_2P = \text{constant}$$

will be a circle. Figures 12.12(b) and (c) show the fringe patterns for $D = 20 \text{ cm}$ and $D = 10 \text{ cm}$; for both figures $S_1S_2 = d = 0.05 \text{ mm}$ and $\lambda = 5000 \text{ \AA}$. Obviously, if O represents the centre of the fringe pattern then

$$S_1O - S_2O = d = 100 \lambda$$

Thus (for this value of d) the central spot will be bright for all values of D and will correspond to $n = 100$. The first and second bright circles will correspond to a path difference of 99λ and 98λ respectively. Similarly, the first and second dark rings in the interference pattern will correspond to a

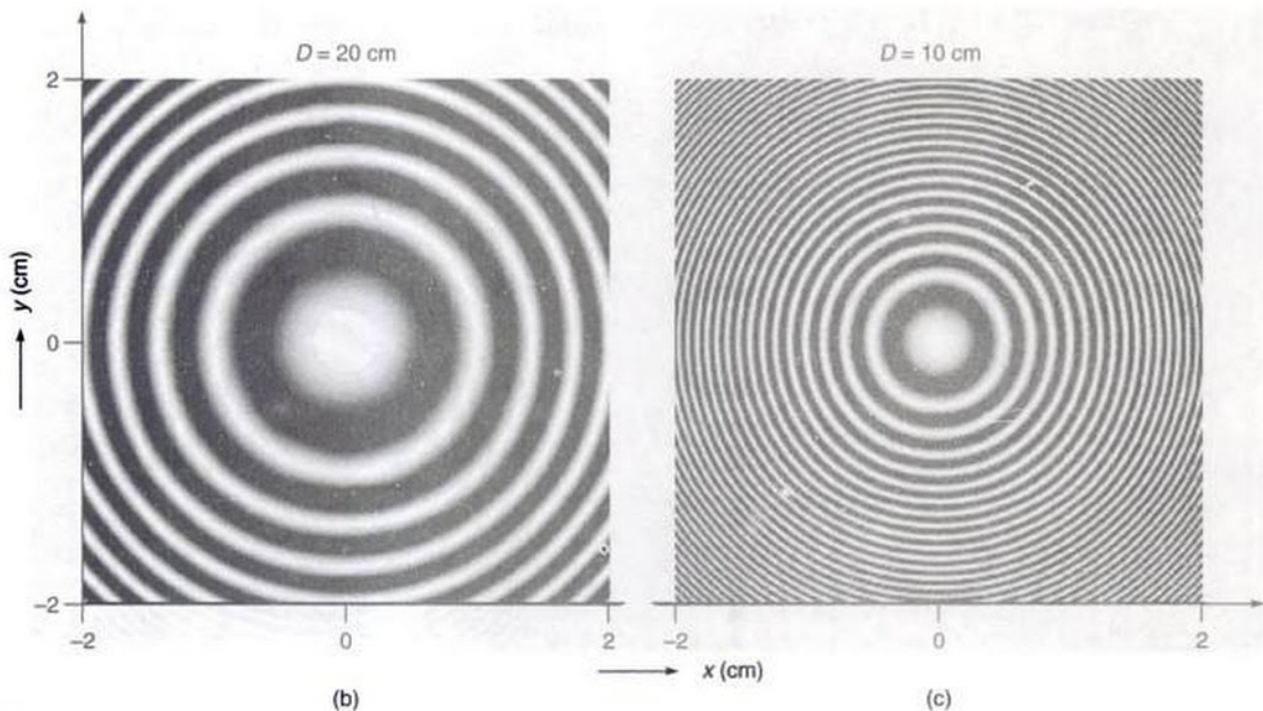
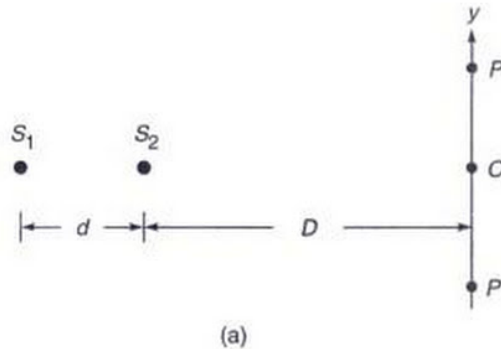


Fig. 12.12 (a) S_1 and S_2 represent two coherent sources, (b) and (c) show the interference fringes observed on the screen PP' when $D = 20 \text{ cm}$ and $D = 10 \text{ cm}$ respectively.

path difference of 99.5λ and 98.5λ respectively. The radii of the fringes can be calculated by using the formula given in Problem 12.10.

Example 12.7 We finally consider the interference pattern produced on PP' by the superposition of a plane wave incident normally and a spherical wave emanating from the point O . The plane wave will be given by

$$E_1 = E_0 \cos(kz - \omega t + \phi)$$

and the spherical wave will be given by

$$E_2 = \frac{A_0}{r} \cos(kr - \omega t)$$

where r is the distance measured from the point O which is assumed to be the origin. Now, on the plane PP' ($z = D$)

$$r = (x^2 + y^2 + D^2)^{\frac{1}{2}} \approx D \left[1 + \frac{x^2 + y^2}{2D^2} \right]$$

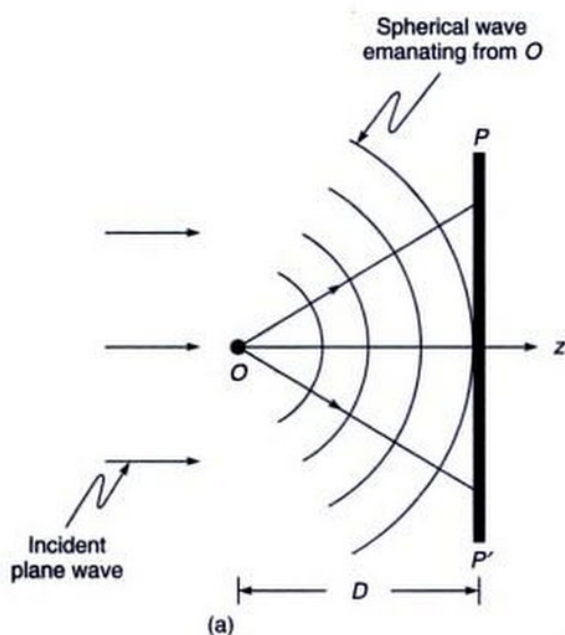
$$\approx D + \frac{x^2 + y^2}{2D}$$

where we have assumed $x, y \ll D$. On the plane $z = D$, the resultant field will be given by

$$E = E_1 + E_2$$

$$\approx E_0 \cos(kD - \omega t + \phi)$$

$$+ \frac{A_0}{D} \cos \left[kD + \frac{k}{2D} (x^2 + y^2) - \omega t \right]$$



Thus

$$\langle E^2 \rangle = \frac{1}{2} E_0^2 + \frac{1}{2} \left(\frac{A_0}{D} \right)^2$$

$$+ E_0 \frac{A_0}{D} \cos \left[\frac{k}{2D} (x^2 + y^2) - \phi \right]$$

If we assume that

$$\frac{A_0}{D} \approx E_0$$

i.e., the amplitude of the spherical wave (on the plane PP') is the same as the amplitude of the plane wave then

$$\langle E^2 \rangle = 2E_0^2 \cos^2 \left[\frac{k}{4D} (x^2 + y^2) - \frac{1}{2} \phi \right]$$

and we would obtain circular interference fringes as shown in Fig. 12.13(b). If r_m and r_{m+p} denote the radii of m^{th} and $(m+p)^{\text{th}}$ bright ring then

$$r_{m+p}^2 - r_m^2 = 2p\lambda D$$

12.6.1 Moire Fringes

We may mention here that Moire fringes can be very effectively used to study the formation of fringe patterns. In Fig. 12.14 we have shown the overlapping of two simple patterns from which one can understand the formation of bright and dark fringes when two plane waves propagate in slightly different directions. In a classroom, it can be easily

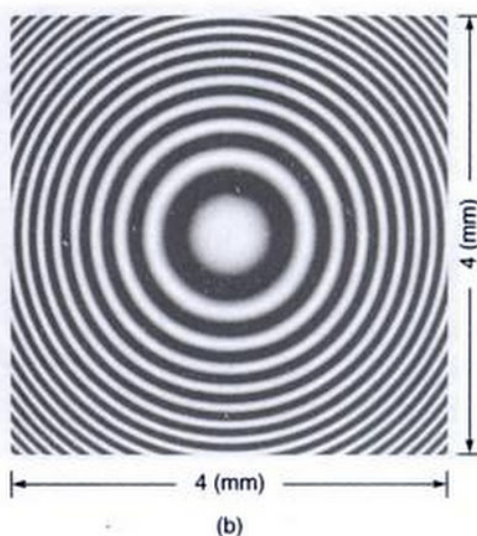


Fig. 12.13 (a) Superposition of a plane wave and a spherical wave emanating from the point O ; (b) shows the interference fringes observed on the screen PP' .

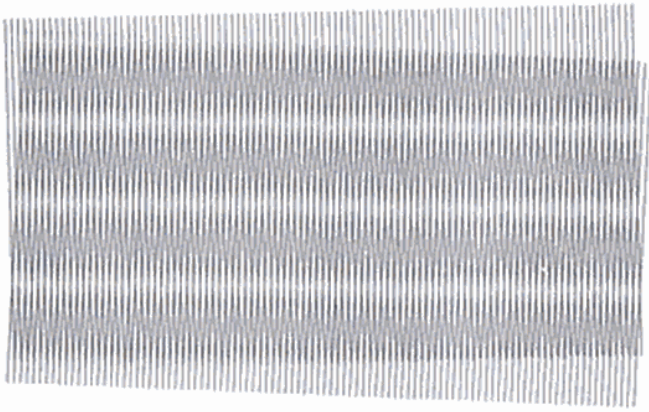


Fig. 12.14 The Moiré Pattern produced by two overlapping straight line patterns.

demonstrated by having a periodic pattern on a transparency and overlapping it with its own photocopy at different angles. Similarly, if one overlaps a circular pattern (on a transparency) with its own copy, one obtains the hyperbolic fringes as shown in Fig. 12.15. (To get a clearer fringe pattern, you may have to view the patterns from a greater distance.) In Sec. 15.5 we have shown how the beat phenomenon can be understood by observing the Moiré fringes obtained by the overlapping of two patterns of slightly different periods (see Fig. 15.13).

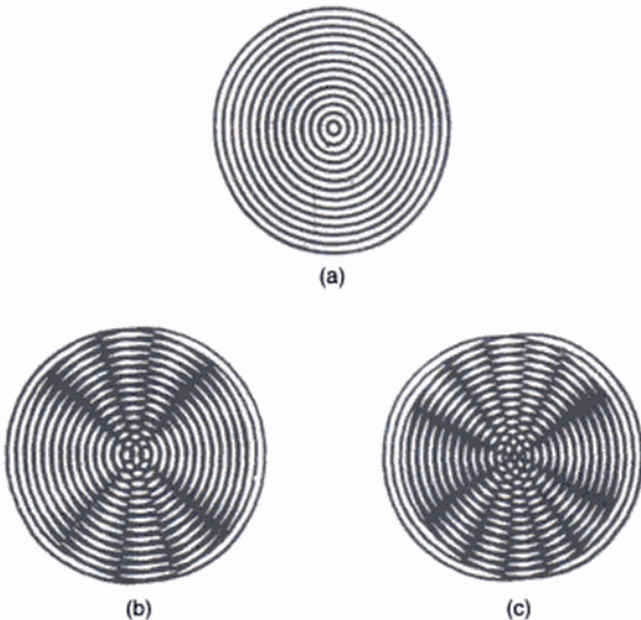


Fig. 12.15 The Moiré pattern produced by two overlapping circular patterns. You will see clear hyperbolic fringes if you put the pattern at a greater distance from the eye. The circular pattern was provided by Dr. R. E. Bailey.

Example 12.8 Consider two parallel slits S_1 and S_2 (perpendicular to the plane of the paper) illuminated by two distant incoherent sources S' and S'' as shown in Fig. 12.16. The angles $S'AO'$ and $S''AO'$ are 10° each. Assuming that both the sources emit almost monochromatic light of the same wavelength λ , determine the intensity pattern on the screen which is at a distance D from the slits (see Fig. 12.16).

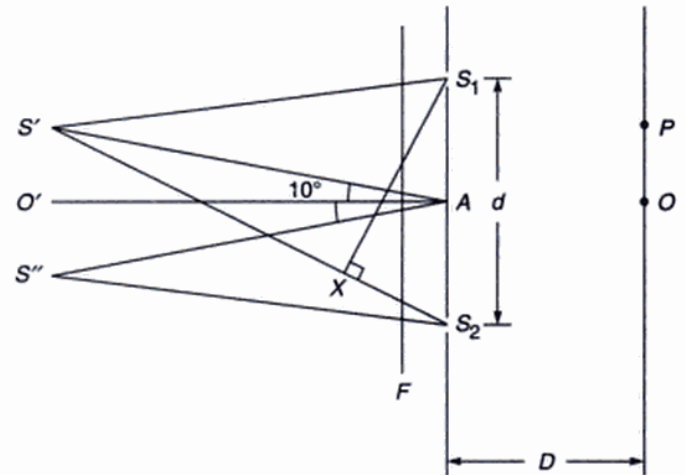


Fig. 12.16 Two distant incoherent sources S' and S'' illuminate the slits S_1 and S_2 .

Solution: We will first calculate the intensity pattern produced by each source. Since the two sources are incoherent, the resultant intensity pattern on the screen will be the sum of the intensities produced by each source on the screen. Consider an arbitrary point P on the screen. Light from the source S' will produce constructive interference at the point P if

$$XS_2 + S_2P - S_1P = n\lambda$$

In the above equation we have assumed S' to be far away from the slits so that $S'S_2 - S'S_1 \approx XS_2$, where X is the foot of the perpendicular drawn from S_1 on $S'S_2$. Thus the intensity pattern on the screen due to S' would be

$$I_{S'} = I_0 \cos^2 \delta/2$$

where

$$\delta = \frac{2\pi}{\lambda} [XS_2 + S_2P - S_1P]$$

But

$$XS_2 = S_1S_2 \sin (XS_1S_2) \approx S_1S_2 \sin 10^\circ$$

Therefore

$$\begin{aligned} I_{S'} &= I_0 \cos^2 \left[\frac{2\pi}{\lambda} (S_2P - S_1P) + \frac{2\pi}{\lambda} S_1S_2 \sin 10^\circ \right] \\ &= I_0 \cos^2 \left[\frac{2\pi}{\lambda} \left\{ \frac{yd}{D} + d \sin 10^\circ \right\} \right] \end{aligned}$$

Similarly

$$I_{S'} = I_0 \cos^2 \left[\frac{2\pi}{\lambda} \left\{ \frac{xd}{D} - d \sin 10^\circ \right\} \right]$$

The resultant intensity would be given by

$$I = I_S + I_{S'}$$

The above example is of practical importance when distant stars (like Betelguse and Rigel) are viewed by means of an interference arrangement. A filter F is usually placed in front of the slits to make the light falling on the slits almost monochromatic.

Example 12.9 In Young's double slit arrangement two calcite half-wave plates* are placed in front of the slit (see Fig. 12.17). The half-wave plate placed in front of slit S_1 has its optic axis in the y -direction, whereas the optic axis of the half-wave plate (placed in front of S_2) is along the x -axis. Determine the intensity distribution on the screen. Assume that the light coming from the source S is polarized along (a) the y -axis and (b) the x -axis.

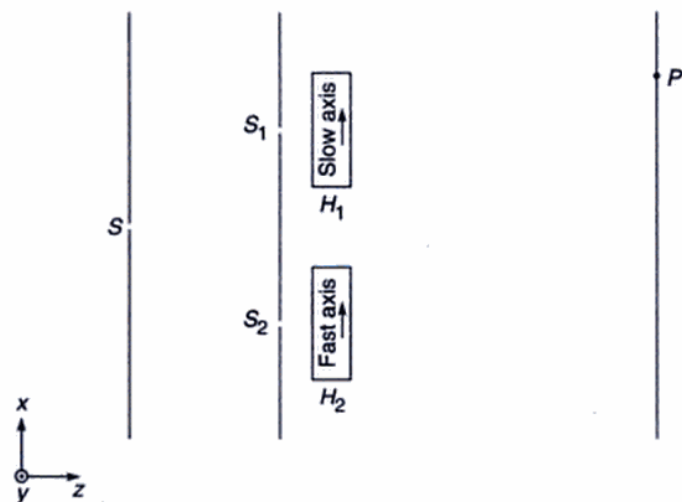


Fig. 12.17 H_1 and H_2 are half-wave plates placed in front of the slits S_1 and S_2 . The optic axis of H_1 and H_2 are along y and x directions respectively.

Solution: Let us first consider the case when \mathbf{E} is along the y -axis. Now, in calcite the extraordinary wave travels faster than the ordinary wave. Furthermore, since the optic axis in H_1 is along the y -axis, a y -polarized wave will propagate as an e -wave through H_1 (see Sec. 19.6). Similarly, a y -polarized wave will propagate as an o -wave through H_2 . Thus, if the path through H_1 contains n wavelengths, then (since the e -wave travels faster) the path through H_2 will

contain $(n + \frac{1}{2})$ wavelengths. Thus there is an additional path of $\lambda/2$ introduced in one of the beams and the entire fringe pattern shifts by half of a fringe, i.e. maxima and minima occur when $S_2P \sim S_1P$ is $(n + \frac{1}{2})\lambda$ and $n\lambda$ respectively. A similar fringe pattern is obtained when \mathbf{E} is in the y -direction. Consequently, the entire fringe pattern would shift by half a fringe.

12.7 FRESNEL'S TWO-MIRROR ARRANGEMENT

After Young's double hole interference experiment, Fresnel devised a series of arrangements to produce the interference pattern. One of the experimental arrangements, known as the Fresnel two-mirror arrangement, is shown in Fig. 12.18; it consists of two plane mirrors which are inclined to each other at a small angle θ and touching at the point M . S represents a narrow slit placed perpendicular to the plane of the paper.

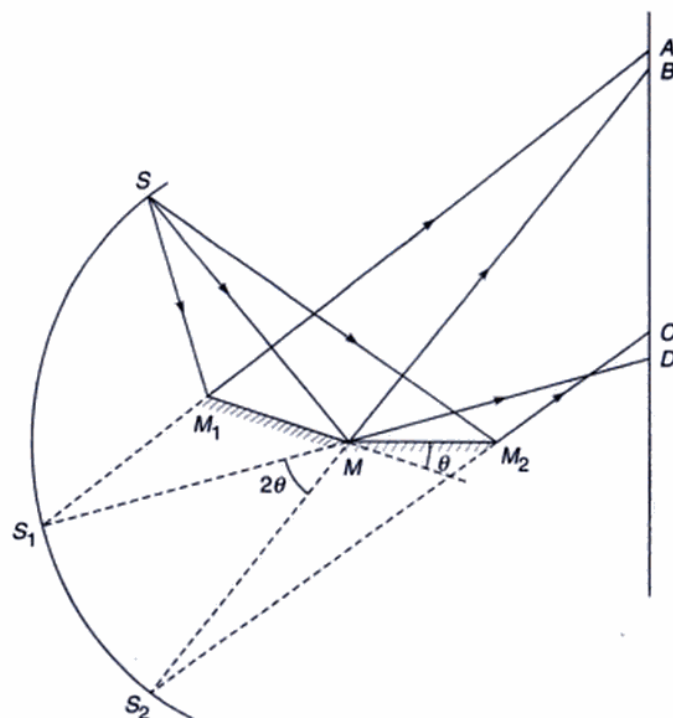


Fig. 12.18 Fresnel's two-mirror arrangement.

A portion of the wavefront from S gets reflected from M_1M and illuminates the region AD of the screen. Another portion of the wavefront gets reflected from the mirror MM_2 and illuminates the region BC of the screen. Since these

*This example presupposes the knowledge of half-wave plates which has been discussed in Chapter 19 [see Sec. 19.6].

two wavefronts are derived from the same source they are coherent. Thus in the region BC , one observes interference fringes. The formation of the fringes can also be understood as being due to the interference of the wavefronts from the virtual sources S_1 and S_2 of S formed by the mirrors M_1 and M_2 respectively. From simple geometric considerations, it can be shown that the points S , S_1 and S_2 lie on a circle whose centre is at the point M . Further, if the angle between the mirrors is θ , then the angle S_1SS_2 is also θ and the angle S_1MS_2 is 2θ . Thus S_1S_2 is $2R\theta$, where R is the radius of the circle.

12.8 FRESNEL BIPRISM

Fresnel devised yet another simple arrangement for the production of interference pattern. He used a biprism, which was actually a simple prism, the base angles of which are extremely small ($\sim 20'$). The base of the prism is shown in Fig. 12.19 and the prism is assumed to stand perpendicular to the plane of the paper. S represents the slit which is also placed perpendicular to the plane of the paper. Light from the slit S gets refracted by the prism and produces two virtual images S_1 and S_2 . These images act as coherent sources and produce interference fringes on the right of the biprism. The fringes can be viewed through an eyepiece. If n represents the refractive index of the material of the biprism and α the base angle, then $(n-1)\alpha$ is approximately the angular deviation produced by the prism and, therefore, the distance S_1S_2 is $2a(n-1)\alpha$, where a represents the distance from S to the base of the prism. Thus, for $n = 1.5$, $\alpha \approx (20') \approx 5.8 \times 10^{-3}$ radians, $a \approx 2$ cm, one gets $d = 0.012$ cm.

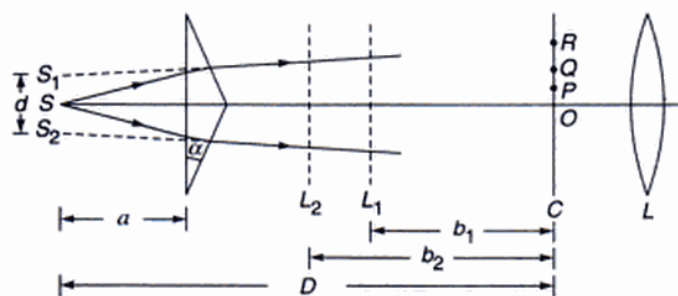


Fig. 12.19 Fresnel's biprism arrangement. C and L represent the positions of the crosswires and the eyepiece respectively. In order to determine d one introduces a lens between the biprism and the crosswires; L_1 and L_2 represent the two positions of the lens where the slits are clearly seen.

The biprism arrangement can be used for the determination of wavelength of an almost monochromatic light like the one coming from a sodium lamp. Light from the sodium lamp illuminates the slit S and interference fringes can be easily viewed through the eyepiece. The fringe width (β) can be determined by means of a micrometer attached to the eyepiece. Once β is known, λ can be determined by using the following relation:

$$\lambda = \frac{d\beta}{D} \quad (34)$$

It may be mentioned that in order to determine d , one need not measure the value of α . In fact the distances d and D can easily be determined by placing a convex lens between the biprism and the eyepiece. For a fixed position of the eyepiece there will be two positions of the lens (shown as L_1 and L_2 in Fig. 12.19) where the images of S_1 and S_2 can be seen at the eyepiece.* Let d_1 be the distance between the two images when the lens is at the position L_1 (at a distance b_1 from the eyepiece). Let d_2 and b_2 be the corresponding distances when the lens is at L_2 . Then, it can easily be shown that

$$d = \sqrt{d_1 d_2}$$

and

$$D = b_1 + b_2$$

Typically for $d \approx 0.01$ cm, $\lambda \approx 6 \times 10^{-5}$ cm, $D \approx 50$ cm, $\beta \approx 0.3$ cm.

In above we have considered here a slit instead of a point source. Since each pair of points S_1 and S_2 produce (approximately) straight line fringes, the slit will also produce straight line fringes of increased intensity.

12.9 INTERFERENCE WITH WHITE LIGHT

We will now discuss the interference pattern when the slit is illuminated by white light. The wavelengths corresponding to the violet and red end of the spectrum are about 4×10^{-5} cm and 7×10^{-5} cm respectively. Clearly, the central fringe produced at the point O (Fig. 12.19) will be white because all wavelengths will constructively interfere here. Now, slightly below (or above) the point O the fringes will become coloured. For example, if the point P is such that

$$S_2P - S_1P = 2 \times 10^{-5} \text{ cm} \left(= \frac{\lambda_{\text{violet}}}{2} \right)$$

then complete destructive interference will occur only for the violet colour. Partial destructive interference will occur

*This method is similar to the displacement method for the determination of the focal length of a convex lens.

for other wavelengths. Consequently we will have a line devoid of the violet colour and will appear reddish. The point Q which satisfies

$$S_2Q \sim S_1Q = 3.5 \times 10^{-5} \text{ cm} \left(= \frac{\lambda_{\text{red}}}{2} \right)$$

will be devoid of the red colour. It will correspond to almost constructive interference for the violet colour. No other wavelength (in the visible region) will neither constructively nor destructively interfere. Thus following the white central fringe we will have coloured fringes; when the path difference is about $2 \times 10^{-5} \text{ cm}$ the fringe will be red in colour, then the colour will gradually change to violet. The coloured fringes will soon disappear because at points far away from O there will be so many wavelengths (in the visible region) which will constructively interfere that we will observe uniform white illumination. For example, at a point R , such that $S_2R \sim S_1R = 30 \times 10^{-5} \text{ cm}$, wavelengths corresponding to $30 \times 10^{-5}/n$ ($n = 1, 2, \dots$) will constructively interfere. In the visible region these wavelengths will be $7.5 \times 10^{-5} \text{ cm}$ (red), $6 \times 10^{-5} \text{ cm}$ (yellow), $5 \times 10^{-5} \text{ cm}$ (greenish yellow) and $4.3 \times 10^{-5} \text{ cm}$ (violet). Further, wavelengths corresponding to $30 \times 10^{-5}/(n + \frac{1}{2})$ will destructively interfere; thus, in the visible region, the wavelengths $6.67 \times 10^{-5} \text{ cm}$ (orange), $5.5 \times 10^{-5} \text{ cm}$ (yellow), $4.6 \times 10^{-5} \text{ cm}$ (indigo) will be absent. The colour of such light, as seen by the unaided eye, will be white. Thus, with white light one gets a white central fringe at the point of zero path difference along with a few coloured fringes on both the sides, the colour soon fading off to white. While using a white light source, if we put a red (or green) filter in front of our eye, we will see the interference pattern corresponding to the red (or green) light.

In the usual interference pattern with a nearly monochromatic source (like a sodium lamp) a large number of interference fringes are obtained and it is extremely difficult to determine the position of the central fringe. In many interference experiments it is necessary to determine the position of the central fringe and, as has been discussed above, this can easily be done by using white light as a source.

12.10 DISPLACEMENT OF FRINGES

We will now discuss the change in the interference pattern produced by introducing a thin transparent plate in the path of one of the two interference beams as shown in Fig. 12.20. Let t be the thickness of the plate and let n be its refractive index. It is easily seen from the figure that light reaching the point P from S_1 has to traverse a distance t in the plate and a distance $S_1P - t$ in air. Thus the time required for the light to reach from S_1 to the point P is given by

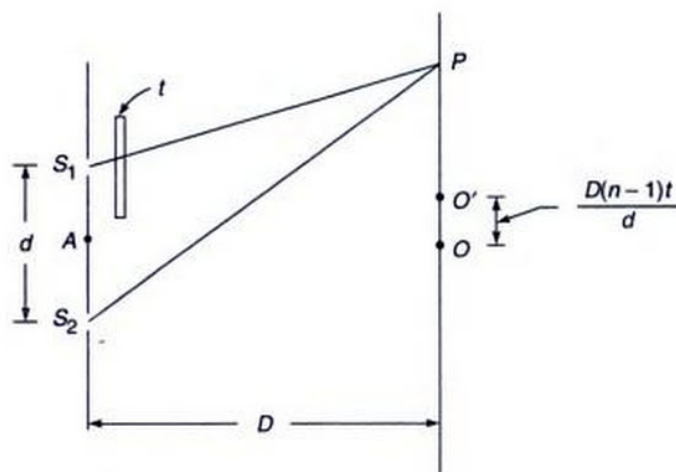


Fig. 12.20 If a thin transparent sheet (of thickness t) is introduced in one of the beams, the fringe pattern gets shifted by a distance $(n-1)tD/d$.

$$\begin{aligned} \frac{S_1P - t}{c} + \frac{t}{v} &= \frac{1}{c} [S_1P - t + nt] \\ &= \frac{1}{c} [S_1P + (n-1)t] \end{aligned} \quad (35)$$

where $v (= \frac{c}{n})$ represents the speed of light in the plate. Eq. (35) shows that by introducing the thin plate the effective optical path increases by $(n-1)t$. Thus, when the thin plate is introduced, the central fringe (which corresponds to equal optical path from S_1 and S_2) is formed at the point O' where

$$S_1O' + (n-1)t = S_2O'$$

Since [see Eq. (19)]

$$S_2O' - S_1O' = \frac{d}{D} OO'$$

therefore

$$(n-1)t = \frac{d}{D} OO' \quad (36)$$

Thus the fringe pattern gets shifted by a distance Δ which is given by the following equation:

$$\Delta = \frac{D(n-1)t}{d} \quad (37)$$

The above principle enables us to determine the thickness of extremely thin transparent sheets (like that of mica) by measuring the displacement of the central fringe. Further, if white light is used as a source, the displacement of the central fringe is easy to measure.

Example 12.10 In a double slit interference arrangement one of the slits is covered by a thin mica sheet whose refractive index is 1.58. The distances S_1S_2 and AO (see Fig. 12.20) are 0.1 cm and 50 cm respectively. Due to the

introduction of the mica sheet the central fringe gets shifted by 0.2 cm. Determine the thickness of the mica sheet.

Solution: $\Delta = 0.2$ cm; $d = 0.1$ cm; $D = 50$ cm
Hence

$$t = \frac{d\Delta}{D(n-1)} = \frac{0.1 \times 0.2}{50 \times 0.58} \\ \approx 6.7 \times 10^{-4} \text{ cm}$$

Example 12.11 In an experimental arrangement similar to the one discussed in the above example one finds that by introducing the mica sheet the central fringe occupies the position that was originally occupied by the eleventh bright fringe. If the source of light is a sodium lamp ($\lambda = 5893 \text{ \AA}$) determine the thickness of the mica sheet.

Solution: The point O' (see Fig. 12.20) corresponds to the eleventh bright fringe, thus

$$S_2O' - S_1O' = 11\lambda = (n-1)t = 0.58t$$

12.11 THE LLOYD'S MIRROR ARRANGEMENT

In this arrangement light from a slit S_1 is allowed to fall on a plane mirror at grazing incidence (see Fig. 12.21). The light directly coming from the slit S_1 interferes with the light reflected from the mirror forming an interference pattern in the region BC of the screen. One may thus consider the slit S_1 and its virtual image S_2 to form two coherent sources which produce the interference pattern. It should be noted that at grazing incidence one really need not have a mirror; even a dielectric surface has very high reflectivity (see Chapter 20).

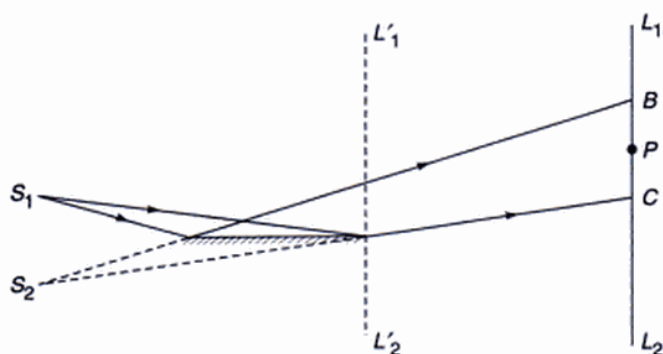


Fig. 12.21 The Lloyd's mirror arrangement.

As can be seen from Fig. 12.21, the central fringe cannot be observed on the screen unless the latter is moved to the position $L'_1L'_2$, where it touches the end of the reflector. Alternatively, one may introduce a thin mica sheet in the path of the direct beam so that the central fringe appears in the region BC . (This is discussed in detail in Problem 12.2) Indeed, if the central fringe is observed with white light, it is found to be dark. This implies that the reflected beam undergoes a sudden phase change of π on reflection. Consequently, when the point P on the screen is such that

$$S_2P - S_1P = n\lambda, n = 0, 1, 2, 3, \dots$$

we will get minima (i.e., destructive interference). On the other hand, if

$$S_2P - S_1P = \left(n + \frac{1}{2}\right)\lambda$$

we will get maxima.

In the next section, using the principle of optical reversibility, we will show that if there is an abrupt phase change of π when light gets reflected by a denser medium, then no such abrupt phase change occurs when reflection takes place at a rarer medium.

12.12 PHASE CHANGE ON REFLECTION

We will now investigate the reflection of light at an interface between two media using the principle of optical reversibility. According to this principle, in the absence of any absorption, a light ray that is reflected or refracted will retrace its original path if its direction is reversed.*

Consider a light ray incident on an interface of two media of refractive indices n_1 and n_2 as shown in Fig. 12.22(a). Let the amplitude reflection and transmission coefficients be r_1 and t_1 respectively. Thus, if the amplitude of the incident ray is a , then the amplitudes of the reflected and refracted rays would be ar_1 and at_1 respectively.

We now reverse the rays and we consider a ray of amplitude at_1 incident on medium 1 and a ray of amplitude ar_1 incident on medium 2 as shown in Fig. 12.22(b). The ray of amplitude at_1 will give rise to a reflected ray of amplitude at_1r_2 and a transmitted ray of amplitude at_1t_2 where r_2 and t_2 are the amplitude reflection and transmission coefficients when a ray is incident from medium 2 on medium 1. Similarly, the ray of amplitude ar_1 will give rise to a ray of amplitude ar_1^2 and a refracted ray of amplitude ar_1t_1 . According to the principle of optical reversibility the two rays

*This principle is consequence of time reversal invariance according to which processes can run either way in time; for more details see Refs. 3 and 8.

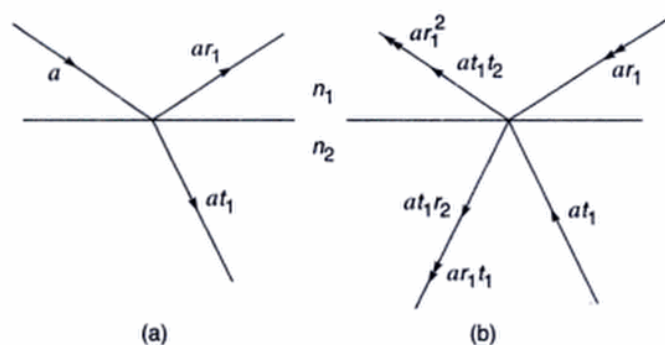


Fig. 12.22 (a) A ray travelling in a medium of refractive index n_1 incident on a medium of refractive index n_2 . (b) Rays of amplitude ar_1 and at_1 incident on a medium of refractive index n_1 .

of amplitudes ar_1^2 and $at_1 t_2$ must combine to give the incident ray of Fig. 12.22(a); thus

$$ar_1^2 + at_1 t_2 = a$$

or

$$t_1 t_2 = 1 - r_1^2 \quad (38)$$

Further, the two rays of amplitudes $at_1 r_2$ and $ar_1 t_1$ must cancel each other, i.e.,

$$at_1 r_2 + ar_1 t_1 = 0$$

or

$$r_2 = -r_1 \quad (39)$$

Since we know from the Lloyd's mirror experiment that an abrupt phase change of π occurs when light gets reflected by a denser medium, we may infer from Eq. (39) that no such abrupt phase change occurs when light gets reflected by a rarer medium. This is indeed borne out by experiments. Equations (38) and (39) are known as Stokes' relations.

In Chapter 21, we will calculate the amplitude reflection and transmission coefficients for plane waves incident on a dielectric and also on a conductor. It will be shown that the coefficients satisfy Stokes' relations; the phase change on reflection will also be discussed there.

SUMMARY

- In 1801, Thomas Young devised an ingenious but simple method to lock the phase relationship between two sources of light. The trick lies in the division of a single wavefront into two; these two split wavefronts act as if they emanated from two sources having a fixed phase relationship and, therefore, when these

two waves were allowed to interfere, a stationary interference pattern is obtained.

- For two coherent point sources, almost straight-line interference fringes are formed on some planes and by measuring the fringe width (which represents the distance between 2 consecutive fringes) one can calculate the wavelength.
- On a plane which is normal to the line joining the two coherent point sources, the fringe pattern is circular.
- In the Young's double slit interference pattern, if we use a white light source, one gets a white central fringe at the point of zero path difference along with a few coloured fringes on both the sides, the colour soon fading off to white. If we now introduce a very thin slice of transparent material (like mica) in the path of one of the interfering beams, the fringes get displaced and by measuring the displacement of fringes, one can calculate the thickness of the mica sheet.

PROBLEMS

- 12.1** In the Young's double-hole experiment (see Fig. 12.6), the distance between the two holes is 0.5 mm, $\lambda = 5 \times 10^{-5}$ cm and $D = 50$ cm. What will be the fringe width?
- 12.2** Figure 12.23 represents the layout of Lloyd's mirror experiment. S is a point source emitting waves of frequency $6 \times 10^{14} \text{ sec}^{-1}$. A and B represent the two ends of a mirror placed horizontally and LOM represents the screen. The distances SP , PA , AB and BO are 1 mm, 5 cm, 5 cm and 190 cm respectively. (a) Determine the position of the region where the

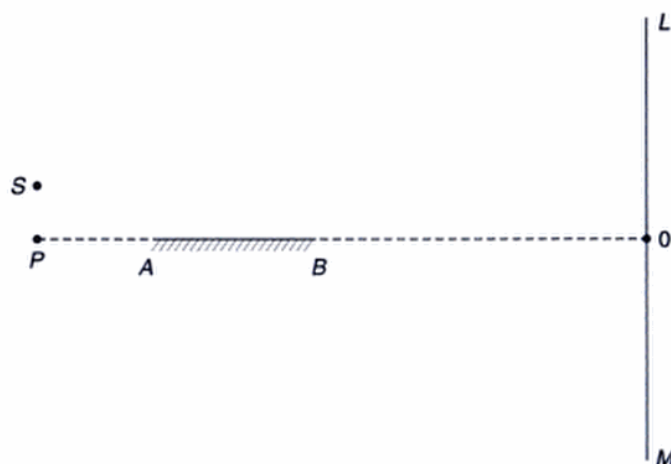


Fig. 12.23 For Problem 12.2.

fringes will be visible and calculate the number of fringes. (b) Calculate the thickness of a mica sheet ($n = 1.5$) which should be introduced in the path of the direct ray so that the lowest fringe becomes the central fringe. The velocity of light is 3×10^{10} cm/sec.

[Ans: (a) 2 cm, 40 fringes, (b) $38 \mu\text{m}$]

- 12.3 (a) In the Fresnel's biprism arrangement, show that $d = 2(n-1)a\alpha$ where a represents the distance from the source to the base of the prism (see Fig. 12.19), α is the angle of the biprism and n the refractive index of the material of the biprism.

(b) In a typical biprism arrangement $b/a = 20$, and for sodium light ($\lambda = 5893 \text{ \AA}$) one obtains a fringe width of 0.1 cm ; here b is the distance between the biprism and the screen. Assuming $n = 1.5$, calculate the angle α .

[Ans: $\approx 0.71^\circ$]

- 12.4 In the Young's double hole experiment a thin mica sheet ($n = 1.5$) is introduced in the path of one of the beams. If the central fringe gets shifted by 0.2 cm , calculate the thickness of the mica sheet. Assume $d = 0.1 \text{ cm}$, and $D = 50 \text{ cm}$.

- 12.5 In order to determine the distance between the slits in the Fresnel biprism experiment, one puts a convex lens in between the biprism and the eye piece. Show that if $D > 4f$ one will obtain two positions of the lens where the image of the slits will be formed at the eye piece; here f is the focal length of the convex lens and D is the distance between the slit and the eye piece. If d_1 and d_2 are the distances between the images (of the slits) as measured by the eye piece, then show that $d = \sqrt{d_1 d_2}$. What would happen if $D < 4f$?

- 12.6 In the Young's double hole experiment, interference fringes are formed using sodium light which predominantly comprises of two wavelengths (5890 \AA and 5896 \AA). Obtain the regions on the screen where the fringe pattern will disappear. You may assume $d = 0.5 \text{ mm}$ and $D = 100 \text{ cm}$.

- 12.7 If one carries out the Young's double hole interference experiment using microwaves of wavelength 3 cm , discuss the nature of the fringe pattern if $d = 0.1 \text{ cm}$, 1 cm , and 4 cm . You may assume $D = 100 \text{ cm}$. Can you use Eq. (21) for the fringe width?

- 12.8 In the Fresnel's two mirror arrangement (see Fig. 12.18) show that the points S , S_1 , and S_2 lie on a circle and $S_1 S_2 = 2b\theta$ where $b = MS$ and θ is the angle between the mirrors.

- 12.9 In the double hole experiment using white light, consider two points on the screen, one corresponding to a path difference of 5000 \AA and the other corresponding to a path difference of 40000 \AA . Find the wavelengths (in the visible region) which correspond to constructive and destructive interference. What will be the colour of these points?

- 12.10 (a) Consider a plane which is normal to the line joining two point coherent sources S_1 and S_2 as shown in Fig. 12.12. If $S_1 P - S_2 P = \Delta$, then show that

$$y = \frac{1}{2\Delta} (d^2 - \Delta^2)^{\frac{1}{2}} [4D^2 + 4Dd + (d^2 - \Delta^2)]^{\frac{1}{2}} \\ \approx \frac{D}{\Delta} \sqrt{(d - \Delta)(d + \Delta)}$$

where the last expression is valid for $D \gg d$.

- (b) For $\lambda = 0.5 \mu\text{m}$, $d = 0.4 \text{ mm}$ and $D = 20 \text{ cm}$; $S_1 O - S_2 O = 800 \lambda$. Calculate the value of $S_1 P - S_2 P$ for the point P to be first dark ring and first bright ring.

[Ans: 0.39975 mm , 0.3995 mm]

- 12.11 In continuation of the above problem calculate the radii of the first two dark rings for

- (a) $D = 20 \text{ cm}$ and (b) $D = 10 \text{ cm}$.

[Ans: (a) $\approx 0.71 \text{ cm}$ and 1.22 cm]

- 12.12 In continuation of Problem 12.10 assume that $d = 0.5 \text{ mm}$, $\lambda = 5 \times 10^{-5} \text{ cm}$ and $D = 100 \text{ cm}$. Thus the central (bright) spot will correspond to $n = 1000$. Calculate the radii of the first, second and third bright rings which will correspond to $n = 999$, 998 and $n = 997$ respectively.

- 12.13 Using the expressions for the amplitude reflection and transmission coefficients (derived in Chapter 21), show that they satisfy Stokes' relations.

- 12.14 Assume a plane wave incident normally on a plane containing two holes separated by a distance d . If we place a convex lens behind the slits, show that the fringe width, as observed on the focal plane of the lens, will be $f\lambda/d$ where f is the focal length of the lens.

- 12.15 In Problem 12.14, show that if the plane (containing the holes) lies in the front focal plane of the lens, then the interference pattern will consist of exactly parallel straight lines. However, if the plane does not lie on the front focal plane, the fringe pattern will be hyperbolae.

- 12.16 In the Young's double hole experiment calculate I/I_{\max} where I represents the intensity at a point where the path difference is $\lambda/5$.

REFERENCES AND SUGGESTED READINGS

1. F. Graham Smith and T.A. King, *Optics and Photonics: An Introduction*, John Wiley, Chichester (2000).
2. R. W. Ditchburn, *Light*, Academic Press, London (1976).
3. R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures in Physics*, Vol. I, Chapter 52, Addison-Wesley, 1965.
4. M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, Cambridge, 2000.
5. E. Hecht and A. Zajac, *Optics*, Addison-Wesley, Reading, Mass., 1974.
6. R. S. Longhurst, *Geometrical and Physical Optics*, 2nd Edition, Longman, London, 1973.
7. D. E. Bailey and M. J. Welch, 'Moire Fringes', *Proceedings of the Conference and Workshop on the Teaching of Optics* (Edited by: G. I. Opat, D. Booth, A. P. Mazzolini and G. Smith, University of Melbourne, Australia).
8. A. Baker, *Modern Physics and Anti Physics*, Chapter 3, Addison-Wesley, Reading, Mass., 1970.
9. PSSC, *Physics*, D.C. Heath & Co. Boston, Mass., 1965.

Chapter 13

Interference by Division of Amplitude

Following a method suggested by Fizeau in 1868, Professor Michelson has produced what is perhaps the most ingenious and sensational instrument in the service of astronomy—the interferometer.

—Sir James Jeans in The Universe Around Us, Cambridge University Press, (1930)

Important Milestones

- 1665 In his treatise, *Micrographia*, the British physicist Robert Hooke described his observations with a compound microscope having a converging objective lens and a converging eye lens. In the same work, he described his observations of the colours produced in flakes of mica, soap bubbles and films of oil on water. He recognised that the colour produced in mica flakes is related to their thickness but was unable to establish any definite relationship between thickness and colour. Hooke advocated a wave theory for the propagation of light
- 1704 "Newton's rings" were first observed by Boyle and Hooke—they are named after Newton because he had given an explanation using the corpuscular model which was later found to be unsatisfactory.
- 1802 Thomas Young gave a satisfactory explanation of 'Newton's rings' based on wave theory.
- 1881 A.A. Michelson invented the "Michelson interferometer". He was awarded the 1907 Nobel Prize in Physics "for his optical precision instruments and the spectroscopic and metrological investigations carried out with their aid" Michelson was America's first Nobel prize winner in science and during the presentation ceremony of the Nobel prize, the President of the Royal Swedish Academy of Sciences said, "Professor Michelson, Your interferometer has rendered it possible to obtain a non-material standard of length possessed of a degree of accuracy never hitherto attained. By its means we are enabled to ensure that the prototype of the meter has remained unaltered in length, and to restore it with absolute infallibility, supposing it were to get lost....."
- 1887 A.A. Michelson and E.W. Morley carried out the famous Michelson-Morley experiment using the Michelson interferometer to detect the motion of the earth with respect to the 'Luminiferous Aether'.

13.1 INTRODUCTION

In the previous chapter we discussed the interference pattern produced by division of a wavefront; for example, light coming out of a pinhole was allowed to fall on two holes, and spherical waves emanating from these two holes produced the interference pattern. In this chapter we will consider the formation of interference pattern by division of amplitude; for example, if a plane wave falls on a thin film then the wave reflected from the upper surface interferes with the wave reflected from the lower surface. Such studies have many practical applications and also explain phenomena like the formation of beautiful colours produced by a soap film illuminated by white light.

13.2 INTERFERENCE BY A PLANE PARALLEL FILM WHEN ILLUMINATED BY A PLANE WAVE

If a plane wave is incident normally on a thin* film of uniform thickness d (see Fig. 13.1) then the waves reflected from the upper surface interfere with the waves reflected from the lower surface; in this section we will study this interference pattern. In order to observe the interference pattern without obstructing the incident beam, we use a partially reflecting plate G as shown in Fig. 13.1. Such an arrangement also enables us to eliminate the direct beam from reaching the photographic plate P (or the eye). The plane wave may be produced by placing an illuminated

*Why the film should be thin is explained in Sec. 13.7.

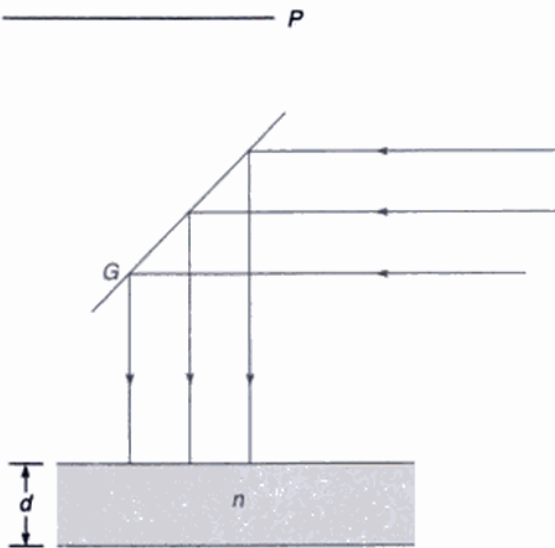


Fig. 13.1 The normal incidence of a parallel beam of light on a thin film of refractive index n and thickness d . G denotes a partially reflecting plate and P represents a photographic plate.

pinhole at the focal point of a corrected lens; alternatively, it may just be a beam coming out of a laser.

Let the solid and the dashed lines in Fig. 13.2 represent the positions of the crests* (at any particular instant of time)



Fig. 13.2 The solid and the dashed lines represent the crests of the waves reflected from the upper surface and from the lower surface of the thin film. Notice that the distance between the consecutive crests inside the film is less than the corresponding distance in medium I.

*Notice that the distance between consecutive crests in the film is less than the corresponding distance in air. This is because of the fact that the effective wavelength in a medium of refractive index n is λ/n .

**In general, the wave reflected from the lower surface of the film will suffer multiple reflections. The effect of such multiple reflections is neglected (see Chapter 14).

corresponding to the waves reflected from the upper and lower surfaces of the film respectively.** Clearly, the wave reflected from the lower surface of the film traverses an additional optical path of $2nd$, where n represents the refractive index of the material of the film. Further, if the film is placed in air, then the wave reflected from the upper surface of the film will undergo a sudden change in phase of π (see Sec. 12.12) and as such the conditions for destructive or constructive interference will be given by

$$2nd = m\lambda \quad \text{destructive interference} \quad (1a)$$

$$= \left(m + \frac{1}{2}\right)\lambda \quad \text{constructive interference} \quad (1b)$$

where $m = 0, 1, 2, \dots$ and λ represents the free space wavelength.

Thus, if we place a photographic plate at P (see Fig. 13.1), then the plate will receive uniform illumination; it will be dark when $2nd = m\lambda$ and bright when $2nd = \left(m + \frac{1}{2}\right)\lambda$; $m = 0, 1, 2, \dots$. Instead of placing the photographic plate, if we try to view the film (from the top) with naked eye, then the film will appear to be uniformly illuminated.

It may be noted that the amplitudes of the waves reflected from the upper and lower surfaces will, in general, be slightly different; and as such the interference will not be completely destructive. However, with appropriate choice of the refractive indices of media II and III, the two amplitudes can be made very nearly equal (see Example 13.1).

For an air film between two glass plates (see Fig. 13.3) no phase change will occur on reflection at the glass-air interface, but a phase change of π will occur on reflection at the air-glass interface and the conditions for maxima and minima will remain the same. On the other hand, if the I medium is crown glass ($n = 1.52$), the II medium is an oil of refractive index 1.60 and the III medium is flint glass ($n = 1.66$) then a phase change of π will occur at both the reflections and the conditions for maxima and minima would be

$$2nd = \left(m + \frac{1}{2}\right)\lambda \quad \text{minima} \quad (2a)$$

$$= m\lambda \quad \text{maxima} \quad (2b)$$

In general, whenever the refractive index of the II medium lies in between the refractive indices of the I and the III media, then the conditions of maxima and minima would be given by Eqs (2a) and (2b).

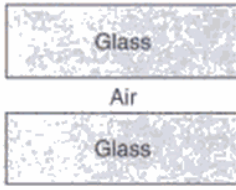


Fig. 13.3 Thin film of air formed between two glass plates.

We next consider the oblique incidence of the plane wave on the thin film (see Fig. 13.4). Once again, the wave reflected from the upper surface of the film interferes with the wave reflected from the lower surface of the film. The latter traverses an additional optical path Δ , which is given by (see Fig. 13.5):

$$\Delta = n_2(BD + DF) - n_1BC \quad (3)$$

where C is the foot of the perpendicular from the point F on BG . We will show in the next section that

$$\Delta = 2n_2d \cos \theta' \quad (4)$$

where θ' is the angle of refraction.

For a film placed in air, a phase change of π will occur when reflection takes place at the point B and as such, the conditions of destructive and constructive interference would be given by

$$\Delta = 2n_2d \cos \theta' = m\lambda \quad \text{minima} \quad (5a)$$

$$= \left(m + \frac{1}{2}\right)\lambda \quad \text{maxima} \quad (5b)$$

If we place a photographic plate at P (see Fig. 13.5) it will receive uniform illumination; if we try to view the film with naked eye (at the position E —see Fig. 13.4) then only light rays reflected from a small position QR of the film will

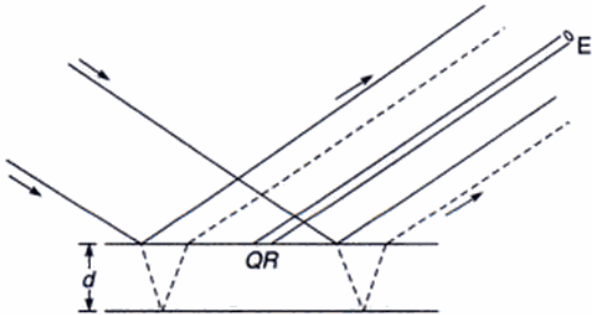


Fig. 13.4 The oblique incidence of a plane wave on a thin film. The solid and dashed lines denote the boundary of the wave reflected from the upper surface and from the lower surface of the film. The eye E receives the light reflected from the region QR .

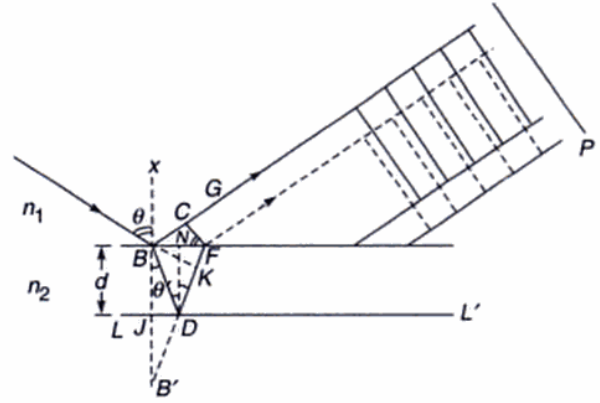


Fig. 13.5 Calculation of the optical path difference between the waves reflected from the upper surface of the film and from the lower surface of the film. The solid and the dashed lines represent the corresponding positions of the crests. P denotes a photographic plate.

reach the eye. The image formed at the retina will be dark or bright depending on the value of Δ (see Eq. 5).

13.3 THE COSINE LAW

In this section we will show that the wave reflected from the lower surface of the film traverses an additional optical path which is given by the following expression:

$$\Delta [= n_2(BD + DF) - n_1BC] = 2n_2d \cos \theta' \quad (6)$$

Let θ and θ' denote the angles of incidence and refraction respectively. We drop a perpendicular BJ from the point B on the lower surface LL' and extend BJ and FD to the point B' where they meet (see Fig. 13.5). Clearly,

$$\angle JBD = \angle BDN = \angle NDF = \theta'$$

where N is the foot of the perpendicular drawn from the point D on BF . Now

$$\angle BDJ = \frac{\pi}{2} - \theta'$$

$$\text{and } \angle B'DJ = \pi - \left[\left(\frac{\pi}{2} - \theta' \right) + \theta' + \theta' \right] = \frac{\pi}{2} - \theta'$$

$$\text{Thus } BD = BD' \text{ and } BJ = JB' = d$$

$$\text{or } BD + DF = B'D + DF = B'F$$

$$\text{Hence } \Delta = n_2B'F - n_1BC \quad (7)$$

$$\text{Now } \angle CFB = \angle CBX = \theta$$

$$BC = BF \sin \theta = \frac{KF}{\sin \theta'} \sin \theta = \frac{n_2}{n_1} KF \quad (8)$$

where K is the foot of the perpendicular from B on $B'F$. Substituting the above expression for BC in Eq. (7) we get

$$\Delta = n_2 B'F - n_2 KF = n_2 B'K$$

$$\text{or} \quad \Delta = 2n_2 d \cos \theta' \quad (9)$$

which is known as the *cosine law*.

13.4 NON-REFLECTING FILMS

One of the important applications of the thin film interference phenomenon discussed in Sec. 13.2 lies in reducing the reflectivity of lens surfaces; this we plan to discuss in this section. However, for a quantitative understanding of the phenomenon, we will have to assume that when a light beam (propagating in a medium of refractive index n_1) is incident normally on a dielectric of refractive index n_2 then the amplitudes of the reflected and the transmitted beams are related to that of the incident beam through the following relations* [See Fig. 13.6(a):

$$a_r = \frac{n_1 - n_2}{n_1 + n_2} a_i \quad (10a)$$

$$a_t = \frac{2n_1}{n_1 + n_2} a_i \quad (10b)$$

where a_i , a_r and a_t are the amplitudes of the incident beam, reflected beam and the transmitted beam respectively. Notice that when $n_2 > n_1$, a_r is negative showing that when a reflection occurs at a denser medium a phase change of π

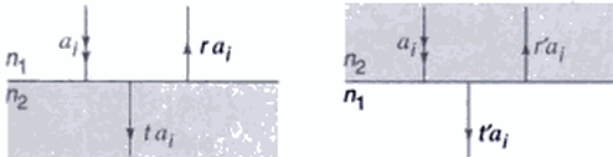


Fig. 13.6 (a) If a plane wave of amplitude a_i , propagating in a medium of refractive index n_1 , is incident normally on a medium of refractive index n_2 , then the amplitudes of the reflected and the transmitted beams are a_r and a_t respectively. Similarly, (b) corresponds to the case when the beam (propagating in a medium of refractive index n_2) is incident on a medium of refractive index n_1 .

occurs. The amplitude reflection and transmission coefficients r and t are, therefore, given by

$$r = \frac{n_1 - n_2}{n_1 + n_2} \quad (11a)$$

$$t = \frac{2n_1}{n_1 + n_2} \quad (11b)$$

It is interesting to point out that if r' and t' are the reflection and transmission coefficients where light propagating in a medium of refractive index n_2 is incident on a medium of refractive index n_1 [see Fig. 13.6(b)], then

$$r' = \frac{n_2 - n_1}{n_2 + n_1} = -r \quad (12)$$

$$t' = \frac{2n_2}{n_1 + n_2} \quad (13)$$

and

$$1 - r'^2 = 1 - \frac{4n_1 n_2}{(n_1 + n_2)^2} = \left(\frac{n_1 - n_2}{n_1 + n_2} \right)^2 = r^2 \quad (14)$$

Equations (13) and (14) represent the Stokes' relations (see Sec. 12.12).

We will now discuss the application of the thin film interference phenomenon in reducing the reflectivity of lens surfaces. We all know that in many optical instruments (like a telescope) there are many interfaces and the loss of intensity due to reflections can be severe. For example, for near normal incidence,** the reflectivity of crown glass surface in air is

$$\left(\frac{n - 1}{n + 1} \right)^2 = \left(\frac{1.5 - 1}{1.5 + 1} \right)^2 \approx 0.04,$$

i.e., 4% of the incident light is reflected. For a dense flint glass $n \approx 1.67$ and about 6% of light is reflected. Thus, if we have a large number of surfaces, the losses at the interfaces can be considerable. In order to reduce these losses, lens surfaces are often coated with a $\lambda/4n$ thick 'non-reflecting film'; the refractive index of the film being less than that of the lens. For example, glass ($n = 1.5$) may be coated with an MgF_2 film (see Fig. 13.7) and the film thickness d should be such that***

$$2n_f d = \frac{1}{2} \lambda$$

$$\text{or} \quad d = \frac{\lambda}{4n_f} \quad (15)$$

*These relations can be derived from electromagnetic theory; see Eqs. (67)-(72) of Chapter 21 (with $\theta_1 = \theta_2 = 0$).

**In all what follows in this section, we will assume near normal incidence.

***Since the refractive index of the non-reflecting film is greater than that of air and less than that of the glass, abrupt phase change of π occurs at both the reflections. Consequently, when $2nd \cos \theta' = m\lambda$ there would be constructive interference and when $2nd \cos \theta' = \left(m + \frac{1}{2}\right)\lambda$ there would be destructive interference.

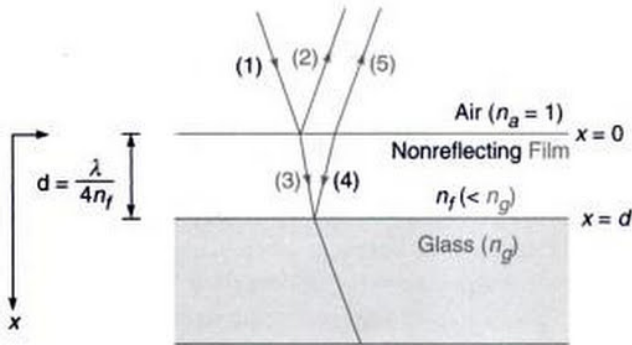


Fig. 13.7 If a film (having a thickness of $\lambda/4n_f$ and having refractive index less than that of the glass) is coated on the glass, then waves reflected from the upper surface of the film destructively interfere with the waves reflected from the lower surface of the film. Such a film is known as a non-reflecting film.

where we have assumed near normal incidence [i.e., $\cos \theta' \approx 1$; see Eq. (9)] and n_f represents the refractive index of the film; for MgF_2 , $n_f = 1.38$. Thus, if we assume λ to be 5.0×10^{-5} cm (which roughly corresponds to the center of the visible spectrum) we will have

$$d = \frac{5.0 \times 10^{-5} \text{ cm}}{4 \times 1.38} \approx 0.9 \times 10^{-5} \text{ cm}$$

We would like to emphasize the following points:

- (a) Let n_a , n_f and n_g be the refractive indices of air, non-reflecting film and glass respectively. If a is the amplitude of the incident wave then the amplitudes of the reflected and refracted waves (the corresponding rays shown as (2) and (3) in Fig. 13.7) would be

$$-\frac{n_f - n_a}{n_f + n_a} a \text{ and } \frac{2n_a}{n_f + n_a} a$$

respectively (we have assumed near normal incidence). The amplitudes of the waves corresponding to rays (4) and (5) would be

$$-\frac{2n_a}{n_f + n_a} \frac{n_g - n_f}{n_g + n_f} a$$

$$\text{and } -\frac{2n_a}{n_f + n_a} \frac{n_g - n_f}{n_g + n_f} \frac{2n_f}{n_f + n_a} a$$

respectively. Now, for complete destructive interference, the waves corresponding to rays (2) and (5) should have the same amplitude, i.e.,

$$-\frac{n_f - n_a}{n_f + n_a} a = -\frac{2n_a}{n_f + n_a} \frac{n_g - n_f}{n_g + n_f} \frac{2n_f}{n_f + n_a} a \quad (16)$$

$$\text{or } \frac{n_f - n_a}{n_f + n_a} = \frac{n_g - n_f}{n_g + n_f} \quad (17)$$

where we have used the fact that $\frac{4n_a n_f}{(n_f + n_a)^2}$ is very nearly equal to unity; for $n_a = 1$ and $n_f = 1.4$,

$$\frac{4n_a n_f}{(n_f + n_a)^2} \approx 0.97$$

On simplification we obtain

$$n_f = \sqrt{n_a n_g} \quad (18)$$

If the first medium is air then $n_a = 1$ and with $n_g = 1.66$ (dense flint glass) n_f should be 1.29 and when $n_g = 1.5$ (light crown glass) n_f should be 1.22. We note that the refractive indices of magnesium fluoride and cryolite are 1.38 and 1.36 respectively. Now for a $\frac{\lambda}{4n}$ thick film, the reflectivity will be about

$$\left[\frac{n_f - n_a}{n_f + n_a} - \frac{n_g - n_f}{n_g + n_f} \right]^2 \quad (19)$$

Thus, for $n_a = 1$, $n_f = 1.38$ and $n_g = 1.5$ the reflectivity will be about 1.3%. In the absence of the film, the reflectivity would have been about 4%. The reduction of reflectivity is much more pronounced for the dense flint glass. This technique of reducing the reflectivity is known as *blooming*.

- (b) The film is non-reflecting only for a particular value of λ ; in Eq. (15) λ was assumed to be 5000 Å. For a polychromatic light, the film's non-reflecting property will be falling off when λ is greater or less than the above value. However, the effect is not serious. For example, for the MgF_2 film on crown glass at 5000 Å, the reflectivity rises by about 0.5% as one goes either to the red or the violet end of the visible spectrum. In Sec. 13.4.2 we will discuss why we should use a $\frac{\lambda}{4n}$ thick film and not $\frac{3\lambda}{4n}$ or $\frac{5\lambda}{4n}$ thick film, although the latter will also give destructive interference for the chosen wavelength.
- (c) As in the case of Young's double slit experiment there is no loss of energy; there is merely a redistribution of energy. The energy appears mostly in the transmitted beam.

13.4.1 Mathematical Expressions for the Reflected Waves

It may be worthwhile to carry out a bit of mathematical analysis for the antireflecting film shown in Fig. 13.7. We

assume $n_g > n_f > n_a$ and that the x -axis is pointing downwards with $x = 0$ at the upper surface of the film. The displacement associated with the incident wave (propagating in the $+x$ direction) is given by

$$y_1 = a \cos(\omega t - k_a x); \quad k_a = \frac{\omega}{c} n_a \quad (20)$$

Thus at $x = 0$, $y_1 = a \cos \omega t$. The reflected wave (shown as 2) would therefore be

$$y_2 = -a |r_1| \cos(\omega t + k_a x); \quad (21)$$

where

$$|r_1| = \left| \frac{n_f - n_a}{n_f + n_a} \right| \quad (22)$$

is a positive quantity. The minus sign in Eq. (21) represents the sudden phase change of π at $x = 0$. The transmitted wave (shown as 3) would be given by

$$y_3 = at_1 \cos(\omega t - k_f x); \quad k_f = \frac{\omega}{c} n_f \quad (23)$$

where

$$t_1 = \frac{2n_a}{n_f + n_a} \quad (24)$$

Thus the displacement at $x = d$ [associated with wave (3)] is

$$y_3 = at_1 \cos(\omega t - k_f d) \quad (25)$$

Therefore, the wave reflected from the lower surface [wave (4), which would be propagating in the negative x -direction] is given by

$$y_4 = -at_1 |r_2| \cos[\omega t + k_f(x - 2d)] \quad (26)$$

$$|r_2| = \left| \frac{n_g - n_f}{n_g + n_f} \right|$$

where the phase factor is adjusted such that at $x = +d$ we obtain the phase given by Eq. (25). The wave (5) would therefore be given by

$$y_5 = -at_1 |r_2| t_2 \cos[\omega t + k_a x - 2k_f d] \quad (27)$$

Assuming the amplitudes of y_2 and y_5 to be approximately the same, destructive interference (between y_2 and y_5) would occur if

$$2k_f d = \pi, 3\pi, \dots \quad (28)$$

or

$$d = \frac{\lambda_f}{4}, \frac{3\lambda_f}{4}, \frac{5\lambda_f}{4}, \dots \quad \lambda_f = \frac{\lambda}{n_f} \quad (29)$$

13.4.2 Rigorous Expressions for Reflectivity

In the above section we have considered two-beam interference and have neglected multiple reflections at the lower and upper surfaces. The effect of multiple reflections will be discussed in Sec. 14.2; however, such an effect is automatically taken into account when we solve Maxwell's equations incorporating the appropriate boundary conditions. In Sec. 21.4 we will carry out such an analysis and will show that the reflectivity (at normal incidence) of a dielectric film of the type shown in Fig. 13.7 is given by [see Eq. (97) of Chapter 21]:*

$$R = \frac{r_1^2 + r_2^2 + 2r_1 r_2 \cos 2\delta}{1 + r_1^2 r_2^2 + 2r_1 r_2 \cos 2\delta} \quad (30)$$

where

$$r_1 = \frac{n_a - n_f}{n_a + n_f} \quad \text{and} \quad r_2 = \frac{n_f - n_g}{n_f + n_g} \quad (31)$$

represent the Fresnel reflection coefficients at the first and second interface respectively, and

$$\delta = \frac{2\pi}{\lambda} n_f d \quad (32)$$

d being the thickness of the film and, as before, λ represents the free space wavelength. Elementary differentiation shows us that $dR/d\delta = 0$ when $\sin 2\delta = 0$. Indeed for $r_1 r_2 > 0$,

$$\cos 2\delta = -1 \quad (\text{minima}) \quad (33)$$

represents the condition for minimum reflectivity and when this condition is satisfied, the reflectivity is given by

$$R = \left(\frac{r_1 - r_2}{1 - r_1 r_2} \right)^2 = \left(\frac{n_a n_g - n_f^2}{n_a n_g + n_f^2} \right)^2 \quad (34)$$

where we have used Eq. (31). Thus the film is non-reflecting when

$$n_f = \sqrt{n_a n_g}$$

consistent with Eq. (18). Now, the condition $\cos 2\delta = -1$ implies

$$2\delta = \frac{4\pi}{\lambda} n_f d = (2m + 1)\pi; \quad m = 0, 1, 2, \dots \quad (35)$$

or

$$d = \frac{\lambda}{4n_f}, \frac{3\lambda}{4n_f}, \frac{5\lambda}{4n_f}, \dots \quad (36)$$

*Equation (30) is actually valid even for oblique incidence with r_1 , r_2 and δ defined appropriately (see Sec. 21.4).

In Fig. 13.8 we have plotted the reflectivity as a function of δ for

$$n_a = 1 \quad n_g = 1.5 \quad (37)$$

and

$$n_f = \sqrt{n_a n_g} \approx 1.225$$

As expected, R is maximum ($\approx 4\%$) when $\delta = 0, \pi, 2\pi, \dots$ and the film is anti-reflecting ($R=0$) when $\delta = \pi/2, 3\pi/2, \dots$ implying $d = \lambda/4n_f, 3\lambda/4n_f, \dots$. As an example, let us suppose that we wish to make the film anti-reflecting at $\lambda = 6000 \text{ \AA}$; then from Eq. (26), the thickness of the film could be

$$1224.7 \text{ \AA} \text{ or } 3674.2 \text{ \AA} \text{ or } 6123.7 \text{ \AA}, \dots$$

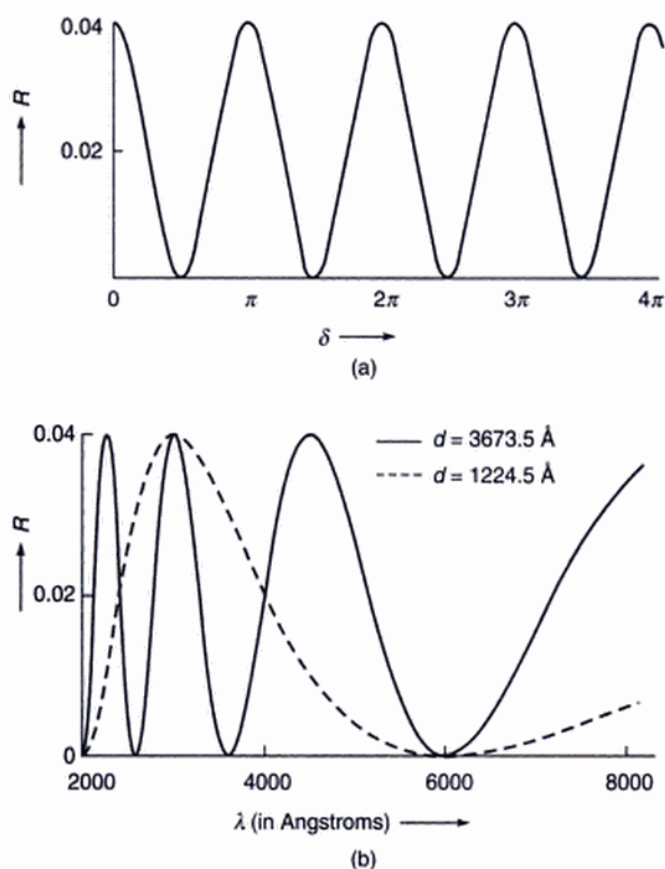


Fig. 13.8 (a) Variation of the reflectivity of a film as a function of $\delta (= 2\pi n_f d/\lambda)$ for $n_a = 1$, $n_g = 1.5$ and $n_f = \sqrt{n_a n_g} \approx 1.225$. Notice that the reflectivity is zero for $\delta = \pi/2, 3\pi/2, 5\pi/2, \dots$ (b) Wavelength variation of the reflectivity for a film of thickness 1224.5 \AA (dashed curve) and of thickness 3673.5 \AA (solid curve) with $n_a = 1$, $n_g = 1.5$ and $n_f = \sqrt{n_a n_g} \approx 1.225$. Notice that both films are anti-reflecting at 6000 \AA .

In Fig. 13.8(b) we have plotted the reflectivity as a function of wavelength for $d = 1224.7 \text{ \AA}$ and 3674.2 \AA . As can be seen, for $d = \lambda/4n_f$, the minimum is broad and the reflectivity small for the entire range of the visible spectrum. Thus for antireflecting coating, the smallest film thickness is always preferred. For $n_a = 1$, $n_g = 1.5$ and $n_f = 1.38$, the reflectivity [according to Eq. (34)] comes out to be 1.4% , which is quite close to the result obtained by using the approximate theory described earlier [see Eq. (19)].

13.5 HIGH REFLECTIVITY BY THIN FILM DEPOSITION

Another important application of the thin film interference phenomenon is the converse of the procedure just discussed, viz., the glass surface is coated by a thin film of suitable material to increase the reflectivity. The film thickness is again $\lambda/4n_f$ where n_f represents the refractive index of the film; however, the film is such that its refractive index is greater than that of the glass; consequently, an abrupt phase change of π occurs only at the air-film interface and the beams reflected from the air-film interface and the film-glass interface constructively interfere. For example, if we consider a film of refractive index 2.37 (zinc sulphide) then the reflectivity is $(2.37 - 1)^2/(2.37 + 1)^2$, i.e., about 16% . In the presence of a glass surface of refractive index 1.5 (light crown glass), the reflectivity will become [see the analysis in Sec. 13.4]:

$$\left[-\frac{2.37 - 1}{2.37 + 1} - \frac{4 \times 1 \times 2.37}{(3.37)^2} \times \frac{2.37 - 1.5}{2.37 + 1.5} \right]^2$$

which gives about 35% . It should be noted that if the difference between the refractive indices of the film and the glass is increased, then the reflectivity will also increase.

We can again use Eq. (30) to calculate the high reflectivity obtained by thin film deposition. Indeed when $n_a < n_f$ and $n_f > n_g$, $r_1 r_2 < 0$ (see Eq. 31) and

$$\cos 2\delta = -1 \text{ (maxima)} \quad (38)$$

represents the condition for *maximum* reflectivity. The *maximum* value of the reflectivity is given by

$$R = \left(\frac{r_1 - r_2}{1 - r_1 r_2} \right)^2 \quad (39)$$

For $n_a = 1.0$, $n_f = 2.37$ and $n_g = 1.5$, we have

$$r_1 \approx -0.407, \quad r_2 \approx 0.225$$

Elementary calculations show that the reflectivity is about 33% which compares well with the value of 35% obtained by using the approximate theory described above.

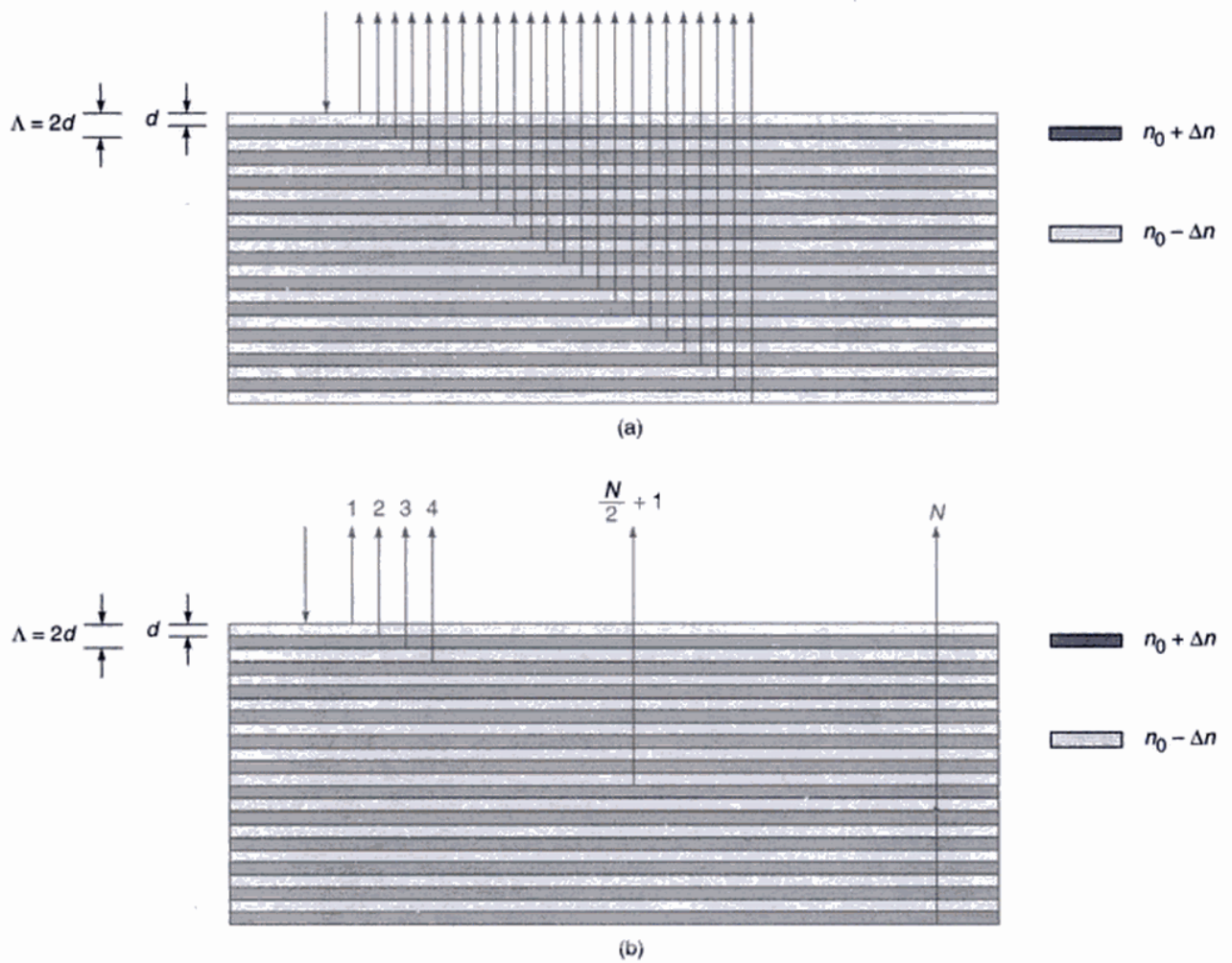


Fig. 13.9 (a) Reflection from a periodic structure consisting of alternate layers of refractive indices ($n_0 + \Delta n$) and ($n_0 - \Delta n$), each of thickness $d = \lambda_B / 4n_0$. (b) If we choose a wavelength ($\lambda_B + \Delta\lambda$) such that reflections from layer 1 and layer ($N/2 + 1$) are out of phase, reflections from layer 2 and layer ($N/2 + 2$) are out of phase etc. and finally reflection from layers $N/2$ and N are out of phase, then the reflectivity will be zero.

13.6 REFLECTION BY A PERIODIC STRUCTURE*

In Sec. 13.4 we had shown that a film of thickness $\lambda/4n_f$ where λ is the free space wavelength and n_f is the film refractive index (which lies between the refractive indices of the two surrounding media) acts like an antireflection layer. This happens due to the destructive interference occurring between the waves reflected from the top and bottom interfaces. In Sec. 13.5 we had shown that if the refractive index of the film was smaller (or greater) than both the surrounding media, then in such a case, in addition to the phase difference due to the additional path travelled by the wave reflected from the lower interface, there would

also be an extra phase difference of π between the two reflected waves. Thus, in such a case a film of thickness $\lambda/4n_f$ would increase the reflectivity rather than reduce it.

We now consider a medium consisting of alternate layers of high and low refractive indices of $n_0 + \Delta n$ and $n_0 - \Delta n$ of equal thickness d [see Fig. 13.9(a)]. Such a medium is called a periodic medium and the spatial period of the refractive index variation is given by

$$\Lambda = 2d$$

Now if $\Delta n \ll n_0$, and if we choose the thickness of each layer to be

$$d = \frac{\lambda}{4n_0} \approx \frac{\lambda}{4(n_0 + \Delta n)} \approx \frac{\lambda}{4(n_0 - \Delta n)}$$

* This section has been very kindly written by Professor K. Thyagarajan.

then the reflections arising out of individual reflections from the various interfaces would all be in phase and should result in a strong reflection. Thus for strong reflection at a chosen (free space) wavelength λ_B , the period of the refractive index variation should be

$$\Lambda = 2d = \frac{\lambda_B}{2n_0} \quad (40)$$

This is referred to as the Bragg condition and is very similar to the Bragg diffraction of X-rays from various atomic layers (see Sec. 16.9). Equation (40) corresponds to the Bragg condition for normal incidence. The quantity λ_B is often referred to as the Bragg wavelength.

As an example, we consider a periodic medium comprising of alternate layers of refractive indices 1.51 and 1.49, i.e., $n_0 = 1.50$ and $\Delta n = 0.01$. If we require a strong reflectivity at $\lambda = \lambda_B = 5500 \text{ \AA}$ then the required periodicity is

$$\Lambda = \frac{5500}{2 \times 1.5} \text{ \AA} \approx 1833 \text{ \AA}$$

If the periodic medium is made up of 100 layers (i.e., 50 periods) then we may approximate the total resultant amplitude to be

$$100 \times \frac{\Delta n}{n_0} \approx \frac{1}{1.5}$$

where $\Delta n/n_0$ is the amplitude reflection coefficient at each interface. The above estimation is only an approximation which is valid when $N \Delta n/n_0 \ll 1$, i.e., for small reflectivities; here we are just trying to obtain a crude estimate of the total reflectivity. Thus the reflectivity at 5500 \AA should be

$$R \approx \left(\frac{1}{1.5} \right)^2 \Rightarrow R \approx 44\% \quad (41)$$

Figure 13.10 shows an actual calculated value of the reflectivity as a function of wavelength (using rigorous electromagnetic theory⁶) for a periodic medium with $n_0 = 1.5$, $\Delta n = 0.01$, $d = \lambda_B / 4n_0$ and consisting of 100 layers. Note that the actual calculation predicts a reflectivity of about 33% which compares well with our crude estimate of 44%!

One notices from Fig. 13.10 that as we move away from the central wavelength ($\lambda_B = 2n_0\Lambda$) the reflectivity of the periodic medium falls off sharply. One can indeed obtain an approximate expression for the wavelength deviation $\Delta\lambda$ from λ_B which will produce a zero reflectivity. In order to do this, we first note that at $\lambda_B (= 2n_0\Lambda)$, the waves reflected from each of the N individual layers are all in phase leading to a strong reflection. If we move away from λ_B then the individual waves reflected from the various layers will not be in phase and thus the reflectivity reduces. If we choose a wavelength $(\lambda_B + \Delta\lambda)$ such that the reflections

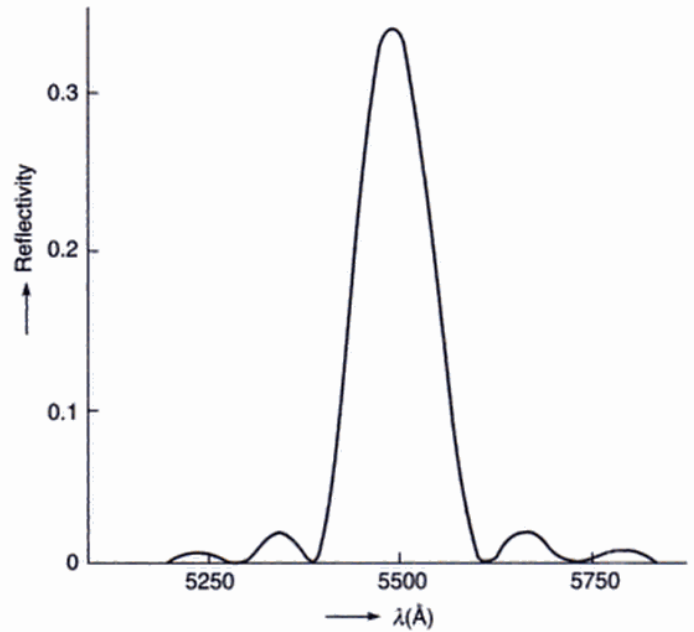


Fig. 13.10 The exact variation of reflectivity with wavelength of a 100 layer periodic structure with $n_0 = 1.5$, $\Delta n = 0.01$, $\Lambda = 2d = 1833 \text{ \AA}$. The peak reflectivity appears at $\lambda = \lambda_B = 4n_0 d$. (Adapted from Ref. 6).

from layer 1 and layer $(\frac{N}{2} + 1)$, from layer 2 and $(\frac{N}{2} + 2)$, and so on up to the reflections from layers $\frac{N}{2}$ and N are out of phase [see Fig. 13.9(b)], then the reflectivity will be zero. For reflection from each of the top $N/2$ layers, there is a reflection from a corresponding lower $N/2$ layer which is out of phase. (The argument is very similar to that used for obtaining the direction of minima in the diffraction pattern of a slit – see Sec. 16.2 and Fig. 16.5). Thus when we move from λ_B to $(\lambda_B + \Delta\lambda)$, the waves reflected from the first and $(\frac{N}{2} + 1)$ th layer should have an additional phase difference of π . Thus,

$$\frac{2\pi}{\lambda_B} n_0 \frac{N\Lambda}{2} - \frac{2\pi}{(\lambda_B + \Delta\lambda)} n_0 \frac{N\Lambda}{2} = \pi \quad (42)$$

where the first term on the LHS is simply the phase difference at λ_B between reflections 1 and $(\frac{N}{2} + 1)$ due to the extra path travelled by the latter wave and the second term is that at $(\lambda_B + \Delta\lambda)$. Assuming $\Delta\lambda \ll \lambda_B$, we have

$$\frac{2\pi}{\lambda_B^2} \frac{n_0 N\Lambda}{2} \Delta\lambda = \pi$$

or

$$\frac{\Delta\lambda}{\lambda_B} = \frac{\lambda_B}{n_0 N\Lambda} = \frac{\Lambda}{L} \quad (43)$$

where we have used Eq. (40) and $L = NA/2$ is the total thickness of the periodic medium. For the example shown in Fig. 13.10, we have

$$\Delta\lambda \approx 110 \text{ \AA} \quad (44)$$

which compares very well with the actual value in Fig. 13.10. Thus if the incident wave is polychromatic (like white light) the reflected light may have a high degree of monochromaticity. This is indeed the principle used in white light holography.

The periodic medium discussed above finds wide applications in high reflectivity multilayer coatings, volume holography, fiber Bragg gratings etc. We will have a very brief discussion on fiber Bragg gratings below.

13.4.1 Fiber Bragg Gratings

A periodic structure discussed above has a very important application in the working of a fiber Bragg grating (usually abbreviated as FBG). We will discuss optical fibers in Chapter 24—it may suffice here to mention that an optical fiber is a cylindrical structure consisting of a central dielectric core cladded by a material of slightly lower refractive index (see Fig. 24.7); the guidance of the light beam takes place because of total internal reflections at the core-cladding interface (see Chapter 24 for details). The cladding material is pure silica and the core is usually silica doped with germanium; the doping results in a slightly higher refractive index. Now, when a germanium-doped silica core fiber is exposed to ultraviolet radiation (with wavelength around $0.24 \mu\text{m}$), the refractive index of the germanium-doped region increases; this is due to the phenomenon known as *photosensitivity* which was discovered by Kenneth Hill in 1974. The refractive index increase can be as large as 0.001 in the core of the fiber. If the fiber is exposed to a pair of interfering UV beams (see

Fig. 13.11), then in regions of constructive interference, the refractive index increases. The two interfering beams are derived from the same beam.

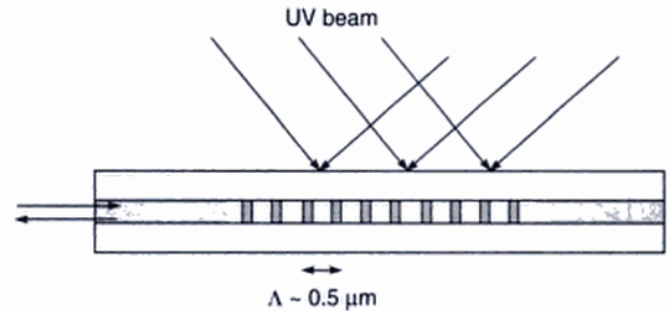


Fig. 13.11 A Fiber Bragg Grating (usually abbreviated as FBG) is produced by allowing two beams to produce an interference pattern.

The period of the grating can be controlled by choosing the angle between the interfering beams. Thus exposing a germanium-doped silica fiber to the interference pattern formed between two UV beams leads to the formation of a periodic refractive index variation in the core of the fiber.

We consider a polychromatic beam incident on the fiber as shown in Fig. 13.11. As discussed above, the reflection from the periodic structure will add up in phase when

$$\Lambda = 2d = \frac{\lambda}{2n_0} \quad (45)$$

which is the Bragg condition. Figure 13.12(a) shows the frequency spectrum of the incident polychromatic beam, the corresponding spectrum of the reflected beam is shown in Fig. 13.12(b). Figure 13.12(c) shows a typical measured frequency spectrum of the reflected wave. For a silica fiber $n_0 \approx 1.46$ and for the periodic structure to be reflecting at $\lambda = 1.55 \mu\text{m}$ we must have

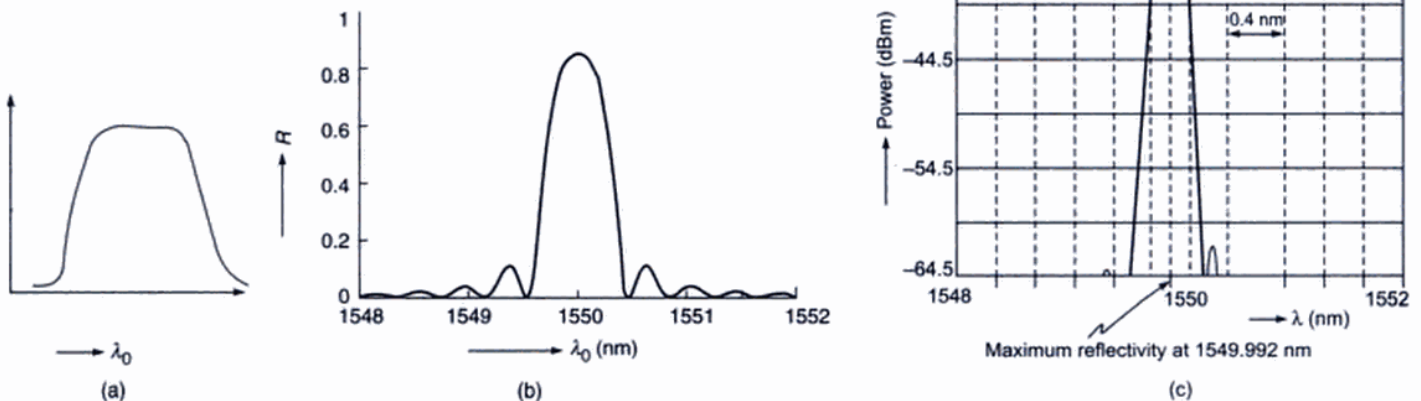


Fig. 13.12 (a) The broad spectrum of the incident wave; (b) the calculated spectrum of the reflected wave which has a very small bandwidth; (c) the measured spectrum of the reflected beam; the center wavelength is 1549.992 nm with a -3dB bandwidth of 0.264 nm. The grating period is 535.48 nm and $n_0 = 1.4474$ [photograph courtesy Mr. Kamal Dasgupta, CGCRI, Kolkata; the FBG was fabricated at CGCRI].

$$\Lambda = \frac{\lambda_B}{2n_0} = \frac{1.55 \mu\text{m}}{2 \times 1.46} \approx 0.531 \mu\text{m} \quad (46)$$

The corresponding peak reflectivity is given by

$$R_p = \tanh^2 \left(\frac{\pi \Delta n L}{\lambda_B} \right) \approx 0.855 \quad (47)$$

where we have assumed $\Delta n = 4 \times 10^{-4}$ and $L = 2 \text{ mm}$. The corresponding bandwidth is given by [cf. Eq. (43)]

$$\frac{\Delta \lambda}{\lambda_B} \approx \frac{\lambda_B}{n_0 L} \left[1 + \left(\frac{(\Delta n) L}{\lambda_B} \right)^2 \right] \quad (48)$$

giving $\Delta \lambda \approx 1 \text{ nm}$. As can be seen from the above equations that the bandwidth (i.e., the monochromaticity of the reflected wave) and the peak reflectivity are determined by Δn and L . Because of the small bandwidth of the reflected spectrum, fiber Bragg gratings find applications in many diverse areas.

13.7 INTERFERENCE BY A PLANE PARALLEL FILM WHEN ILLUMINATED BY A POINT SOURCE

In Sec. 13.2 we had considered the incidence of a parallel beam of light on a thin film and had discussed the interference produced by the waves reflected from the upper and lower surfaces of the film. We will now consider the illumination of the film by a point source of light and, once again, in order to observe the film without obstructing the incident beam, we will use a partially reflecting plate as shown in Fig. 13.13. However, in order to study the interference pattern we may assume the point source S to be right above the film (see Fig. 13.13) such that the distance SK (in Fig. 13.14) is equal to $SA + AK$ (in Fig. 13.13); KA (in Fig. 13.13) and KS (in Fig. 13.14) being normal to the film. Obviously, the waves reflected from the upper surface of the film will appear to emanate from the point S' where

$$KS' = KS \quad (49)$$

Further, simple geometrical considerations will show that the waves reflected from the lower surface will appear to emanate from the point S'' , where

$$KS'' = KS + 2d/n_2 \quad (50)$$

(see Fig. 13.14). Equation (50) is valid only for near normal incidence.* Thus, at least for near normal incidence, the interference pattern produced in region I (see Fig. 13.14)

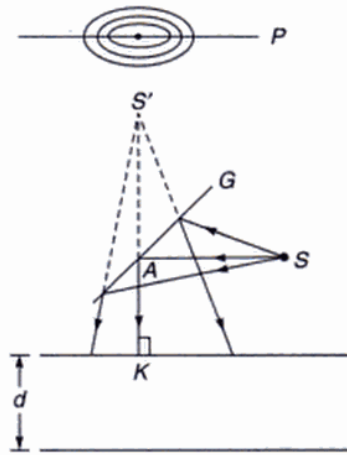


Fig. 13.13 Light emanating from a point source S is allowed to fall on a thin film of thickness d . G is a partially reflecting plate and P represents the photographic plate. On the photographic plate circular fringes are obtained.

will be very nearly** the same as produced by two point coherent sources S' and S'' (which is the double hole experiment of Young discussed in the previous chapter). Thus, if we put a photographic plate P (see Fig. 13.13) we will, in general, obtain interference fringes. The intensity of an arbitrary point Q will be determined by the following relations:

$$\Delta = \left(m + \frac{1}{2} \right) \lambda \quad \text{maxima} \quad (51a)$$

$$= m\lambda \quad \text{minima} \quad (51b)$$

$$\text{where } \Delta = [n_1 SF + n_2 (FG + GH) + n_1 HQ] - [n_1 (SA + AQ)] \quad (52)$$

represents the optical path difference and we have assumed that in one of the reflections, an abrupt phase change of π occurs; n_1 and n_2 are the refractive indices of media I and II respectively. The above conditions are rigorously correct; i.e., valid even for large angles of incidence. Further, it can be shown that for near normal incidence,

$$\Delta \approx 2n_2 d \cos \theta' \quad (53)$$

A more rigorous calculation shows

$$\Delta \approx 2n_2 d \cos \theta' \left[1 - \frac{n_1^2 \sin \theta \cos \theta}{n_2^2 - n_1^2 \sin^2 \theta} \left(\frac{\theta_0 - \theta}{2} \right) \right] \quad (54)$$

where the angles θ , θ_0 and θ' are defined in Fig. 13.14.

Now, if we put a photographic plate (parallel to the surface of the film (see Fig. 13.14)) we will obtain dark and

*This is a consequence of the fact that the image of a point source produced by a plane refracting surface is not perfect.

**The fact that this is not identical to the Young's pattern is because of the fact that S'' is not a perfect image of the point S . For large angles of incidence, the waves reflected from the lower surface will appear to emanate from a point which will be displaced from S'' .

bright concentric rings (see Example 12.6).^{*} On the other hand, if we view the film with naked eye then, for a given position of the eye, we will be able to see only a very small portion of the film; e.g., with eye at the position E and the point source at S only a portion of the film around the point B will be visible [see Fig. 13.15(a)], and this point will appear to be dark or bright as the optical path difference,

$$\Delta = n_1 SQ + n_2(QA + AB) - n_1 SB$$

is $m\lambda$ or $(m + \frac{1}{2})\lambda$. Further, using a method similar to the one described in Sec. 13.3, we can obtain

$$\Delta \approx 2n_2 d \cos \theta' \quad (55)$$

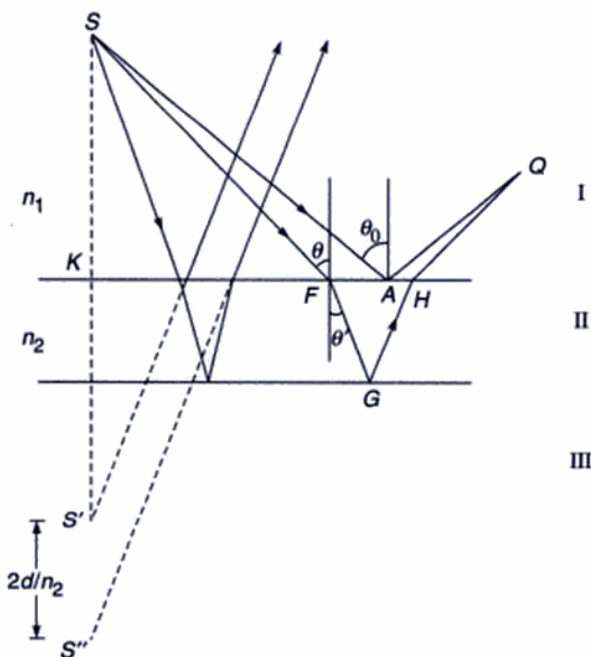


Fig. 13.14 If light emanating from a point source S is incident on a thin film then the interference pattern produced in the region I is approximately the same as would have been produced by two coherent point sources S' and S'' (separated by a distance $2d/n_2$) where d represents the thickness of the film and n_2 represents the refractive index of the film.

Instead of looking at the film, if the eye is focussed at infinity, then the interference is between the rays which are derived from a single incident ray by reflection from the upper and the lower surfaces of the film [see Fig. 13.15(b)]. For example, the rays PM and QR , which focus at the point O of the retina, are derived from the single ray SP , and the rays $P'M'$ and $Q'R'$, which focus at a different point O' on

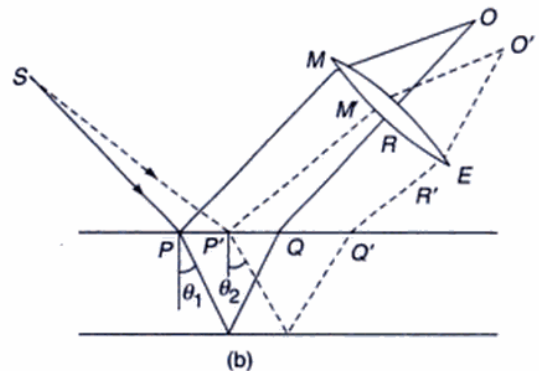
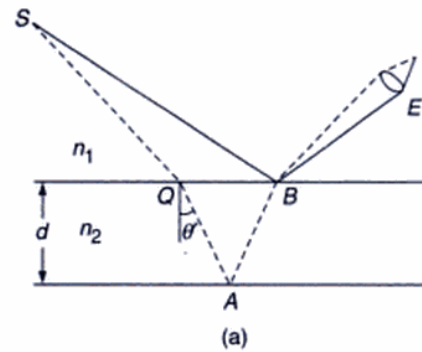


Fig. 13.15 Light emanating from a point source S is incident on a thin film; (a) if the film is viewed by the naked eye E then the point B will appear to be dark if the optical path difference $[(n_1 SQ + n_2(QA + AB)) - n_1 SB]$ is $m\lambda$, and bright if the optical path is $(m + \frac{1}{2})\lambda$. (b) If the eye is focussed for infinity then it receives parallel rays from different directions corresponding to different values of the angles of refraction θ' (and hence different values of the optical path difference).

the retina, are derived from the ray SP' . Since the angles of refraction θ'_1 and θ'_2 (for these two sets of rays) will be different, the points O and O' will, in general, not have the same intensity.

We next consider the illumination by an extended source of light S (see Fig. 13.16). Such an extended source may be produced by illuminating a ground glass plate by a sodium lamp. Each point on the extended source will produce its own interference pattern on the photographic plate P ; these will be displaced with respect to one another; consequently, no definite fringe pattern will appear on the photographic plate. However, if we view the film with our eye, rays from all points of the film will reach the eye. If the eye is focussed at infinity then parallel light coming in a particular direction reaching the eye would have originated from nearby points

^{*}If the point source is taken far away, then it can easily be seen that the rings will spread out and in the limit of the point source being taken to infinity (i.e., incidence of a parallel beam), the photographic plate will be uniformly illuminated.

of the extended source and the intensity produced on the retina would depend on the value of $2nd \cos \theta'$ which is the same for all parallel rays like S_1Q , S_2Q' , etc. (see Fig. 13.16). Rays emanating in a different direction (like S_1R , S_2R' , etc.) would correspond to a different value of θ' and would focus at a different point on the retina. Since θ' is constant over the circumference of a cone (whose axis is normal to the film and whose vertex is at the eye), the eye will see dark and bright concentric rings, with the center lying along the direction $\theta' = 0$. Such fringes, produced by a film of uniform thickness, are known as Haidinger fringes. They are also known as fringes of equal inclination because the changes in the optical path are due to the changes in the direction of incidence and hence in the value of θ' . In Sec. 13.10 we will discuss the Michelson interferometer where such fringes are observed.

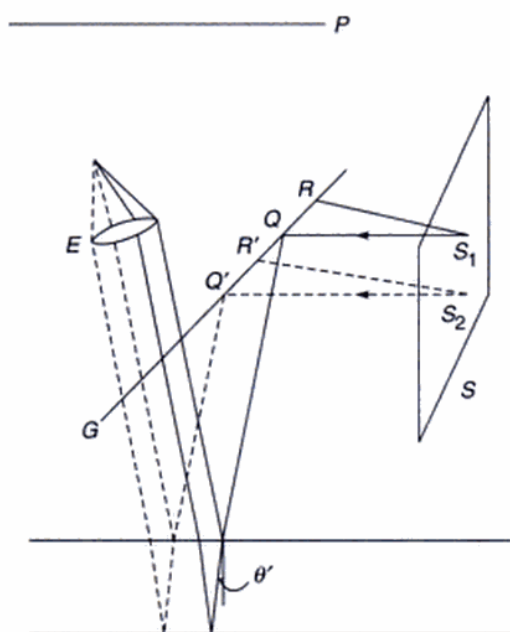
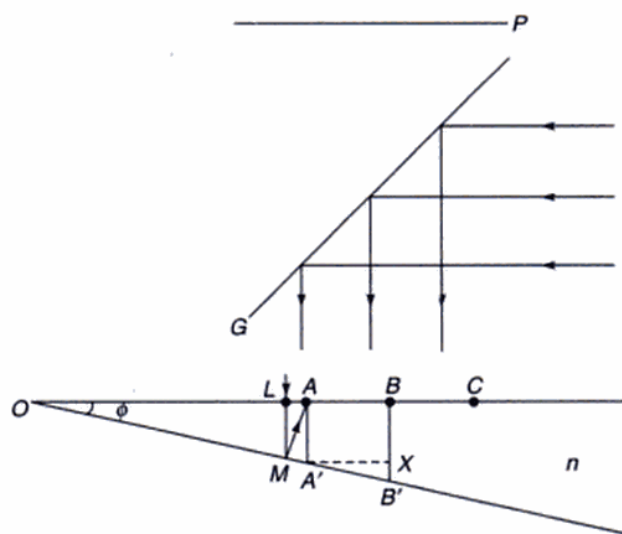


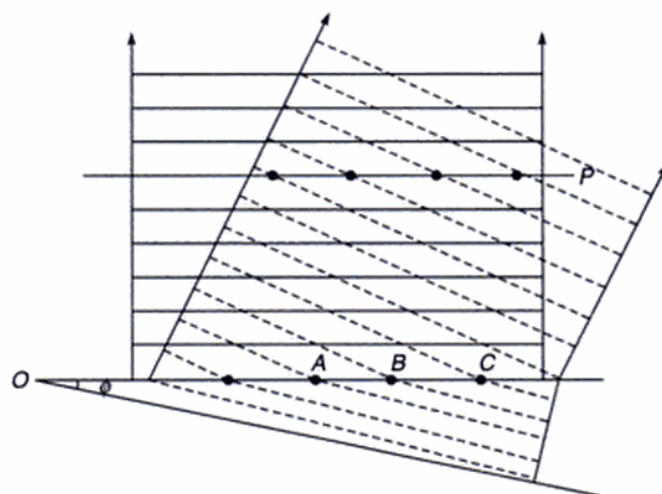
Fig. 13.16 Light emanating from an extended source illuminates a thin film. G represents the partially reflecting plate and P represents the photographic plate. The eye E is focussed at infinity.

13.8 INTERFERENCE BY A FILM WITH TWO NON-PARALLEL REFLECTING SURFACES

Till now we have assumed the film to be of uniform thickness. We will now discuss the interference pattern produced by a film of varying thickness. Such a film may be produced by a wedge which consists of two non-parallel plane surfaces [see Fig. 13.17(a)].



(a)



(b)

Fig. 13.17 (a) A parallel beam of light incident on a wedge. (b) The solid and the dashed lines represent the positions of the crests (at a particular instant of time) corresponding to the waves reflected from the upper surface and from the lower surface respectively. The maxima will correspond to the intersection of the solid and dashed lines. The fringes will be perpendicular to the plane of the paper.

We first consider a parallel beam of light incident normally on the upper surface of the film [see Fig. 13.17(a)]. In Fig. 13.17(b) the successive positions of the crests (at a particular instant of time) reflected from the upper surface and from the lower surface of the film are shown by solid and dashed lines respectively. Obviously, a photographic

plate P will record straight line interference fringes which will be parallel to the edge of the wedge (the edge is the line passing through the point O and perpendicular to the plane of the paper). The dots in the figure indicate the positions of maxima. In order to find the distance between two consecutive fringes on the film we note that for the point A to be bright*

$$n(LM + MA) = \left(m + \frac{1}{2}\right)\lambda; \quad m = 0, 1, 2, \dots \quad (56)$$

[see Fig. 13.17(a)]. However, when the wedge angle ϕ is very small (which is indeed the case for practical systems)

$$LM + MA \approx 2AA'$$

where AA' represents the thickness of the film at A . Thus the condition for the point A to be bright is

$$2nAA' = \left(m + \frac{1}{2}\right)\lambda \quad (57)$$

Similarly, the next bright fringe will occur at the point B where

$$2nBB' = \left(m + \frac{3}{2}\right)\lambda \quad (58)$$

$$\text{Thus } 2n(BB' - AA') \approx \lambda$$

$$\text{or } XB' \approx \lambda/2n \quad (59)$$

$$\text{But } XB' = (A'X) \tan \phi$$

$$\text{or } A'X = \beta \approx \frac{\lambda}{2n\phi} \quad (60)$$

where β represents the fringe width and we have assumed ϕ to be small. Such fringes are commonly referred to as *fringes of equal thickness*.

On the other hand, for a point source, the fringe pattern will be similar to the parallel film case; i.e., for near normal incidence, the pattern will be very nearly the same as produced by two sources S' and S'' (Fig. 13.18). (Notice that the point S'' is not vertically below S' ; this is a consequence of the fact that the two surfaces of the film are not parallel.) The intensity of an arbitrary point Q will be determined by the following equations:

$$\begin{aligned} [SA + n(AB + BC) + CQ] - [SD + DQ] \\ = \left(m + \frac{1}{2}\right)\lambda \quad \text{maxima} \\ = m\lambda \quad \text{minima} \end{aligned} \quad (61)$$

* We are assuming here that the beam undergoes a sudden phase change of π when it gets reflected by the upper surface. The expression for the fringe width (Eq. 60) is, however, independent of this condition.

** There is, however, one exception to this, when the extended source is taken to a very large distance, then the light rays reaching the plate G will be approximately parallel and an interference pattern (of low contrast) will be formed on the plate P . The same phenomenon will also occur if instead of moving the extended source we take the plate P far away from the wedge.

If we view the film with naked eye (say at the position E — see Fig. 13.18) then only a small portion of the film (around the point R) would be visible and the point R will be bright or dark as the optical path difference $[\{SN + n(NL + LR)\} - SR]$ is $(m + \frac{1}{2})\lambda$ or $m\lambda$ respectively. One can similarly discuss the case when the eye is focussed for infinity.

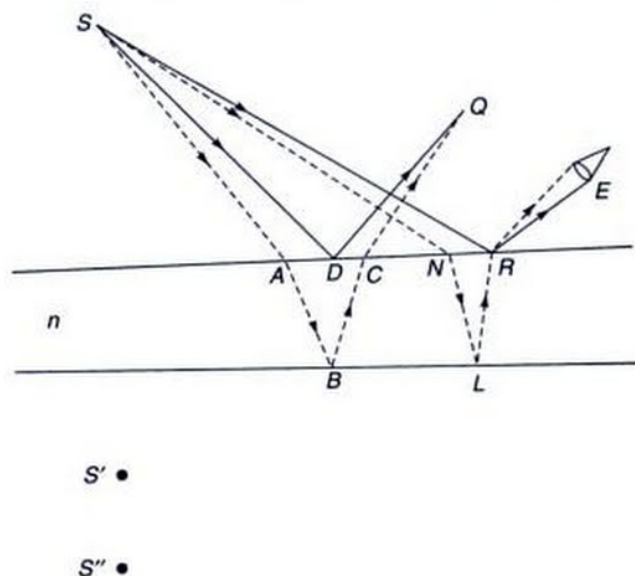


Fig. 13.18 Light from a point source illuminating a wedge. E represents the lens of the eye.

We next consider the illumination by an extended source S as shown in Fig. 13.19. Since the extended source can be assumed to consist of a large number of independent point sources, each point source will produce its own pattern on a photographic plate P ; consequently, no definite fringe pattern will be observed.** However, if we view the film with a camera (or with a naked eye) and if the camera is focussed on the upper surface of the film then a particular point on the film will appear dark or bright depending on the fact that whether $2nd$ is $m\lambda$ or $(m + \frac{1}{2})\lambda$ (see Fig. 13.19) — we are assuming near normal incidence. It may be seen in the figure that interference at the point Q may occur due to light coming from different points on the extended source, but if the incidence is near normal then the intensity at the point Q will be determined entirely by the thickness of the film at that place. Similarly, the intensity at the point Q' will be determined by the thickness of the film at Q' ; however, the point Q' will be focussed at a different point B' on the

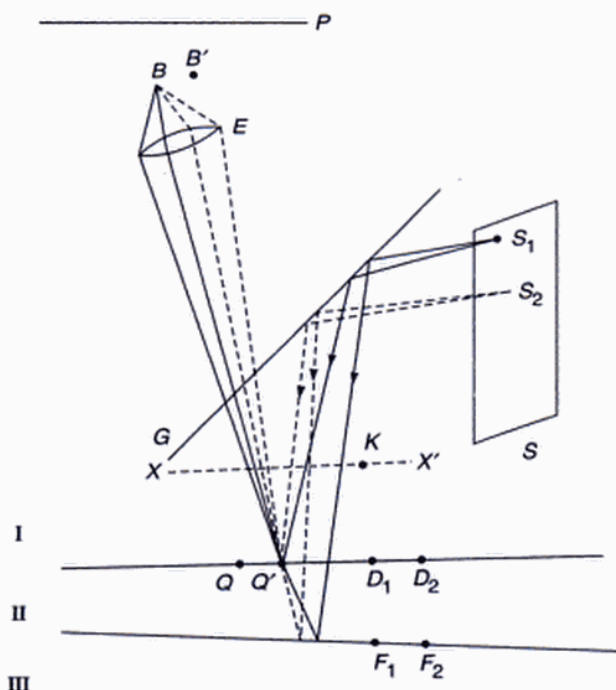


Fig. 13.19 Localized interference fringes produced by an extended source S . Fringes will be seen only when the eye is focussed on the upper surface of the film.

retina of the eye. The fringes will be straight lines parallel to the edge of the film OO' (Fig. 13.20). It should be emphasized that all along we are assuming near normal incidence and the fact that the wedge angle is extremely small. These assumptions are indeed valid for practical systems.

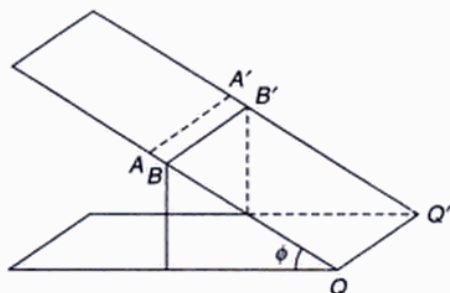


Fig. 13.20 The fringes formed by a wedge will be parallel to the edge OO' .

It is of interest to mention that if we focus the camera on a plane XX' , which is slightly above the film, then no definite interference pattern will be observed. This follows from the fact that the light waves reaching the point K from

S_2 undergo reflection at the points D_2 and F_2 and the light waves reaching K from S_1 undergo reflection at the points D_1 and F_1 . Since the thickness of the film is not uniform, the waves reaching K from S_1 may produce brightness, whereas the waves reaching from S_2 may produce darkness. Thus, in order to view the fringes, one must focus the camera on the upper surface of the film, and in this sense, the fringes are said to be localized. It is left as an exercise for the reader to verify that if the camera is focussed for infinity, no definite interference pattern will be recorded.

Till now we have assumed the film to be 'thin'; the question now arises as to how thin the film should be. In order to obtain an interference pattern, there should be definite phase relationship between the waves reflected from the upper surface of the film and from the lower surface of the film. Thus the path difference $\Delta (= 2nd \cos \theta')$ should be small compared to the coherence length.* For example, if we are using the D_1 line of an ordinary sodium lamp ($\lambda = 5.890 \times 10^{-5}$ cm), the coherence length is of the order of 1 cm and for fringes to be visible Δ should be much less than 1 cm. It should be pointed out that there is no particular value of Δ for which the fringes disappear; but as the value of Δ increases, the contrast of the fringes becomes poorer. A laser beam has a very high coherence length and fringes can be visible even for path differences much greater than 1 m. On the other hand, if we use a white light source no fringes will be visible for $\Delta \geq 2 \times 10^{-4}$ cm (see Sec. 12.9).

It should be pointed out that interference also occurs in region III (see Fig. 13.21) between the directly transmitted beam and the beam which comes out of the film after suffering two reflections, first from the lower surface and then from the upper surface of the film. However, the two amplitudes will be very different and the fringes will have very poor contrast (see Example 13.1).

Example 13.1 Consider a film of refractive index 1.36 in air. Assuming near normal incidence ($\theta \approx 0$), show that whereas the amplitudes of the reflected rays (1) and (5) (Fig. 13.21) are nearly equal, the amplitudes of the transmitted rays (4) and (7) are quite different. (This is the reason why the fringes observed in transmission have very poor contrast.)

Solution: Let the amplitude of the incident ray be a and let the amplitudes of the rays (1), (2), (3),... be denoted by a_1, a_2, \dots etc. Using Eqs (10a) and (10b), we get

$$a_1 = \frac{1-n}{1+n} a = -\frac{0.36}{2.36} a \approx -0.153a$$

*Coherence length is defined in Sec. 15.1. If a source remains coherent for a time τ , then the coherence length (L) will be about $c\tau$, where c is the speed of light in free space. Thus for $\tau_c \sim 10^{-10}$ sec, $L \sim 3$ cm.

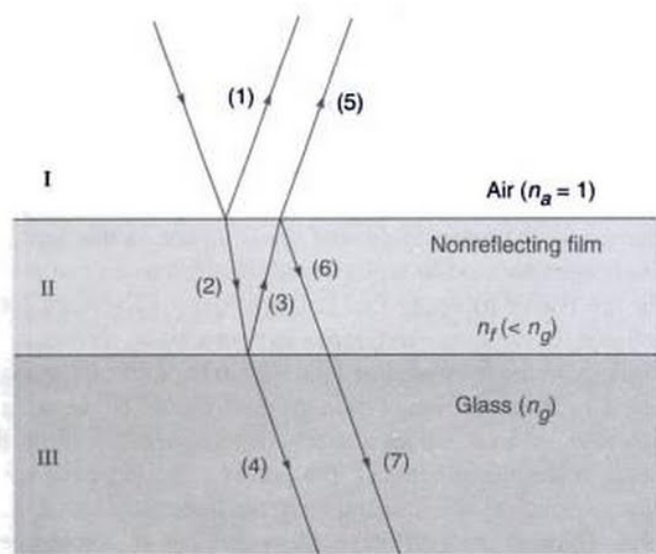


Fig. 13.21 In general, whereas the amplitude of (1) and (5) are nearly the same, the amplitudes of (2) and (6) are quite different.

$$a_2 = \frac{2}{1+n} a = \frac{2}{2.36} a = 0.847a$$

$$a_3 = \frac{n-1}{n+1} a_2 = \frac{0.36}{2.36} \times 0.847a = 0.129a$$

$$a_5 = \frac{2n}{n+1} a_3 = \frac{2 \times 1.36}{2.36} \times 0.129a = 0.149a$$

$$a_4 = \frac{2n}{1+n} a_2 = \frac{2 \times 1.36}{2.36} \times 0.847a = 0.977a$$

$$a_7 = \frac{2n}{n+1} a_6 = \frac{2n}{n+1} \cdot \frac{n-1}{n+1} a_3 = \frac{2 \times 1.36 \times 0.36}{(2.36)^2} a_3 \\ \approx 0.023a$$

We first note that the sign of a_5 is opposite to that of a_1 which is a consequence of the fact that a sudden phase change of π occurs when the ray gets reflected at the point B . Further the magnitude of a_5 is nearly equal to that of a_1 . On the other hand $|a_7| \ll |a_4|$. This is the reason why the interference fringes formed in transmission have poor contrast.

13.9 COLOURS OF THIN FILMS

We have seen in the previous section that if light from an extended monochromatic source (like a sodium lamp) is incident normally on a wedge, then equally spaced dark and bright fringes will be observed. The distance between two consecutive bright (or dark) fringes is determined by the wedge angle, the wavelength of light and by the refractive

index of the film. If we use a polychromatic source (like an incandescent lamp) we will observe coloured fringes. Further, if instead of a wedge we have a film of arbitrarily varying thickness we will again observe fringes, each fringe representing the locus of constant film thickness (see Fig. 13.22). This is indeed what we see when sunlight falls on a soap bubble. It should be mentioned that if the optical path difference between the waves reflected from the upper surface of the film and from the lower surface of the film exceeds a few wavelengths, the interference pattern will be washed out due to the overlapping of interference patterns of many colours and no fringes will be seen (see Sec. 12.9). Thus, in order to see the fringes with white light, the film should not be more than few wavelength thick.



Fig. 13.22 A typical fringe pattern produced by an air-film formed between two glass surfaces (which are not optically flat) and placed in contact with each other. Whenever the thickness of the air-film is $m\lambda/2$, we obtain a dark fringe and when the thickness is $(m + \frac{1}{2})\lambda/2$, we obtain a bright fringe. Each fringe describes a focus of equal thickness of the film. (Photograph courtesy Prof. R.S. Sirohi.)

13.10 NEWTON'S RINGS

If we place a plano-convex lens on a plane glass surface, a thin film of air is formed between the curved surface of the lens (AOB) and the plane glass plate (POQ)—see Fig. 13.23. The thickness of the air film is zero at the point of contact O and increases as one moves away from the point of contact. If we allow monochromatic light (such as from a sodium lamp) to fall on the surface of the lens, then the light reflected from the surface AOB interferes with the light

reflected from the surface POQ . For near normal incidence (and considering points very close to the point of contact) the optical path difference between the two waves is very nearly equal to $2nt$, where n is the refractive index of the film and t the thickness of the film. Thus, whenever the thickness of the air film satisfies the condition

$$2nt = \left(m + \frac{1}{2}\right)\lambda; \quad m = 0, 1, 2, \dots \quad (62)$$

we will have maxima. Similarly the condition

$$2nt = m\lambda \quad (63)$$

will correspond to minima. Since the convex side of the lens is a spherical surface, the thickness of the air film will be constant over a circle (whose centre will be at O) and we will obtain concentric dark and bright rings. These rings are known as Newton's rings.* It should be pointed out that in order to observe the fringes, the microscope (or the eye) has to be focussed on the upper surface of the film (see the discussion in Sec. 13.7).

The radii of various rings can easily be calculated. As mentioned earlier, the thickness of the air film will be constant over a circle whose center is at the point of contact O . Let the radius of the m th dark ring be r_m and if t is the thickness of the air film where the m th dark ring appears to be formed, then

$$r_m^2 = t(2R - t) \quad (64)$$

where R represents the radius of curvature of the convex surface of the lens (see Fig. 13.24). Now $R \approx 100$ cm and $t \leq 10^{-3}$ cm, thus we may neglect t in comparison to $2R$ to obtain

$$r_m^2 \approx 2Rt$$

or

$$2t = \frac{r_m^2}{R} \quad (65)$$

Substituting this in Eq. (63), we get

$$r_m^2 \approx m\lambda R; \quad m = 0, 1, 2, \dots \quad (66)$$

which implies that the radii of the rings vary as square root of natural numbers. Thus the rings will become close to each other as the radius increases (see Fig. 13.25). Between the two dark rings there will be a bright ring whose radius will be $\sqrt{m + \frac{1}{2}} \lambda R$.

Newton's rings can easily be observed in the laboratory by using an apparatus as shown in Fig. 13.23. Light from an extended source (emitting almost monochromatic light, like a sodium lamp) is allowed to fall on a glass plate which

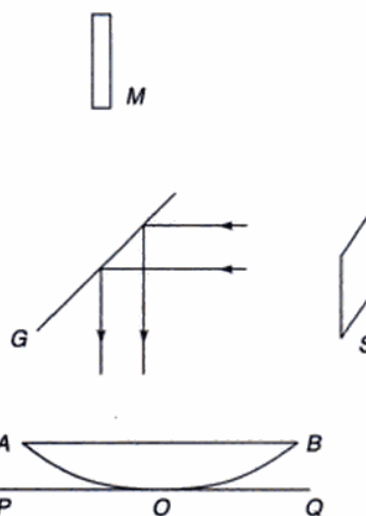


Fig. 13.23 An arrangement for observing Newton's rings. Light from an extended source S is allowed to fall on a thin film formed between the planoconvex lens AOB and the plane glass plate POQ . M represents a travelling microscope.

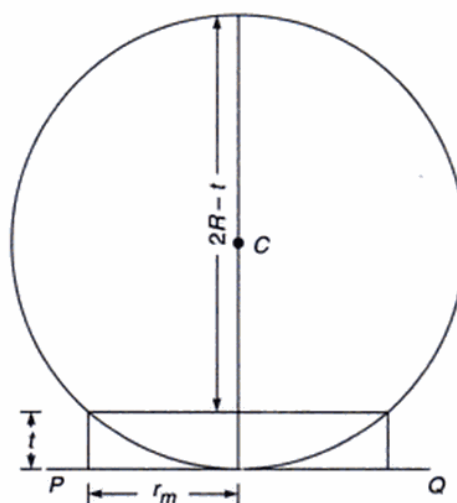


Fig. 13.24 r_m represents the radius of the m th dark ring; the thickness of the air film (where the m th dark ring is formed) is t .

partially reflects the beam. This reflected beam falls on the plano-convex lens-glass plate arrangement and Newton's rings can easily be observed by viewing directly or through a travelling microscope M . Actually, one really need not have plano-convex lens; the rings would be visible even when a biconvex lens is used.

*Boyle and Hooke had independently observed the fringes earlier but Newton was the first to measure their radii and make an analysis. The proper explanation was given by Thomas Young. Also see 'Milestones' in the beginning of this chapter.

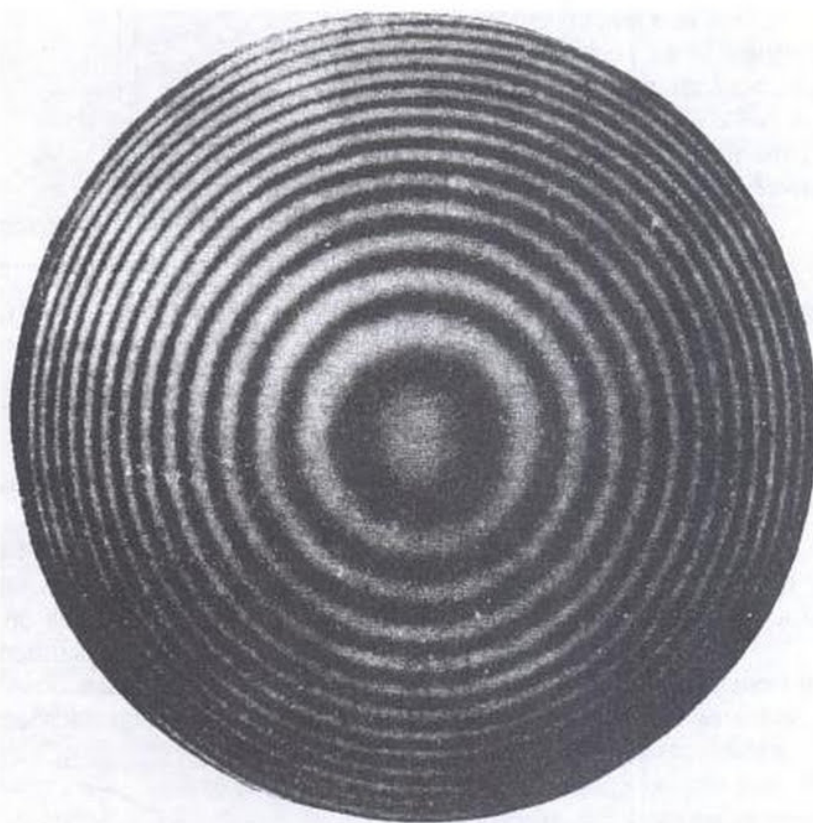


Fig. 13.25 Newton's rings as observed in reflection. The rings observed with transmitted light are of much poorer contrast. (Photograph courtesy Dr. G. Bose.)

Typically for $\lambda = 6 \times 10^{-5}$ cm, and $R = 100$ cm

$$r_m = 0.0774 \sqrt{m} \text{ cm} \quad (67)$$

Thus the radii of the first, second and third dark rings would be approximately 0.0774 cm, 0.110 cm and 0.134 cm respectively. Notice that the spacing between the second and third dark rings is smaller than the spacing between the first and second dark rings.

Equation (63) predicts that the central spot should be dark. Normally, with the presence of minute dust particles the point of contact is really not perfect and the central spot may not be perfectly dark. Thus while carrying out the experiment one should measure the radii of the m th and the $(m+p)$ th ring ($p \approx 10$) and take the difference in the squares of the radii ($r_{m+p}^2 - r_m^2 = p\lambda R$), which is indeed independent of m . Usually, the diameter can be more accurately measured and in terms of the diameters the wavelength is given by the following expression:

$$\lambda = \frac{D_{m+p}^2 - D_m^2}{4pR} \quad (68)$$

The radius of curvature can be accurately measured with the help of a spherometer and therefore by carefully

measuring the diameters of dark (or bright) rings one can experimentally determine the wavelength.

It may be mentioned that if a liquid of refractive index n is introduced between the lens and the glass plate, the radii of the dark rings would be given by

$$r_m = (m\lambda R/n)^{1/2} \quad (69)$$

Equation (69) may be compared with Eq. (66). Further, if the refractive indices of the material of the lens and of the glass plate are different and if the refractive index of the liquid lies in between the two values, the central spot will be bright as in Fig. 13.25 and Eq. (69) would give the radii of the bright rings.

An important practical application of the principle involved in the Newton's rings experiment lies in the determination of the optical flatness of a glass plate. Consider a glass surface placed on another surface whose flatness is known. If a monochromatic light beam is allowed to fall on this combination and the reflected light is viewed by a microscope, then, in general, dark and bright patches will be seen (Fig. 13.22). The space between the two glass surfaces forms an air film of varying thickness and whenever this thickness becomes $m\lambda/2$, we see a dark spot and when this

thickness becomes $(m + \frac{1}{2}) \lambda/2$ we see a bright spot. Two consecutive dark fringes will be separated by the air film whose thickness will differ by $\lambda/2$. Consequently, by measuring the distance between consecutive dark and bright fringes one can calculate the optical flatness of a glass plate.

Example 13.2 Consider the formation of Newton's rings by monochromatic light of $\lambda = 6.4 \times 10^{-5}$ cm. Assume the point of contact to be perfect. Now slowly raise the lens vertically above the plate. As the lens moves gradually away from the plate, discuss the ring pattern as seen through the microscope. Assume the radius of the convex surface to be 100 cm.

Solution: Since the point of contact is perfect, the central spot will be dark, the first dark ring will form at P where $PA = \lambda/2$, and the radius of this ring, OA , will be $\sqrt{\lambda R}$ ($= 0.080$ cm)—see Fig. 13.26(a). Similarly, the radius of the second dark ring will be $OB = \sqrt{2\lambda R}$ ($= 0.113$ cm). If we now raise the lens by $\frac{\lambda}{4}$ ($= 1.6 \times 10^{-5}$ cm) then $2t$ corresponding to the central spot would be $\lambda/2$ and instead of the dark spot at the centre we will now have a bright spot. The radii of the first and the second dark rings will be

$$OA_1 = \left(\frac{1}{2} \lambda R\right)^{1/2} = 0.0566 \text{ cm}$$

$$\text{and } OB_1 = \left(\frac{3}{2} \lambda R\right)^{1/2} = 0.098 \text{ cm}$$

respectively [see Fig. 13.26(b)]. If the lens is further moved by $\lambda/4$ (see Fig. 13.26(c)), then the first dark ring collapses to the centre and the central spot will be dark. The ring which was originally at Q now shifts to Q_2 ; similarly the ring at R [Fig. 13.26(a)] collapses to R_2 [Fig. 13.26(c)].

Thus, as the lens is moved upward the rings collapse to the centre. Hence if we can measure the distance by which the lens is moved upward and also count the number of dark spots that have collapsed to the centre, we can determine the wavelength. For example, in the present case, if the lens is moved by 6.4×10^{-3} cm, 200 rings will collapse to the center. If one carries out this experiment it will be observed that the 200th dark ring will slowly converge to the center and when the lens has moved exactly by 6.4×10^{-3} cm it has exactly come to the center.

Example 13.3 Consider the formation of Newton's rings when two closely spaced wavelengths are present; for example, the D_1 and D_2 lines of sodium ($\lambda_1 = 5890 \text{ \AA}$ and

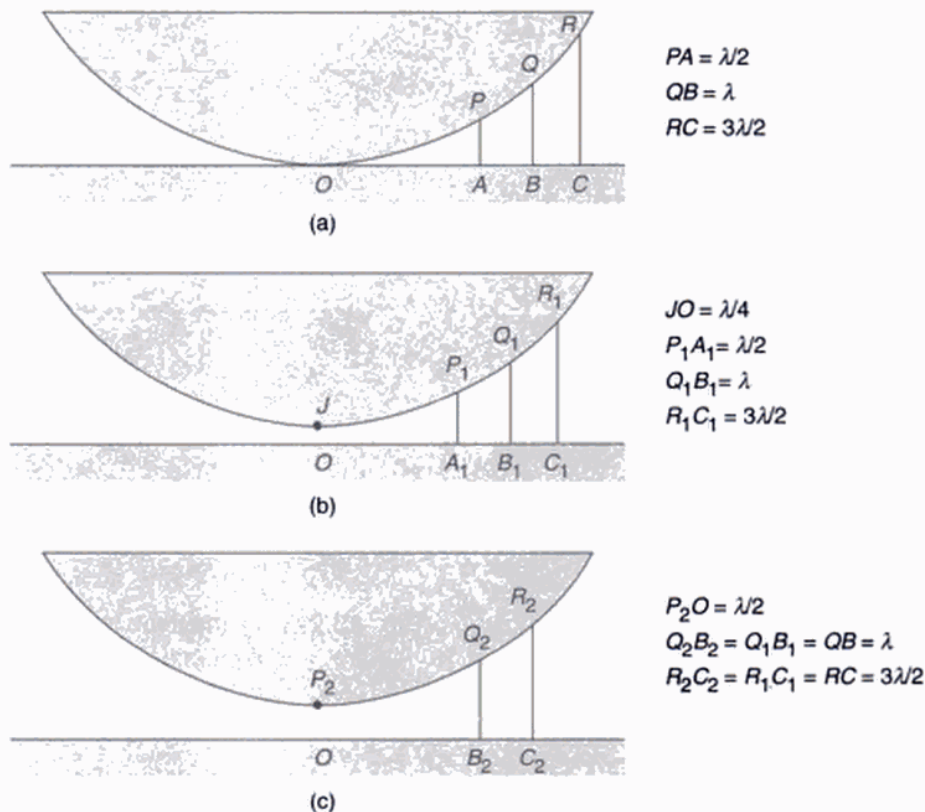


Fig. 13.26 The rings collapse to the centre as the lens is moved away from the plate.

$\lambda_2 = 5896 \text{ \AA}$). What will be the effect of the presence of these two wavelengths as the lens is gradually moved away from the plate? What will happen if the sodium lamp is replaced by a white light source?

Solution: We will first assume that the lens is in contact with the plane glass plate [see Fig. 13.26(a)]. Since the two wavelengths are very close, the bright and dark rings of λ_1 superpose on the bright and dark rings of λ_2 respectively. This can easily be seen by calculating the radii of the ninth dark and bright ring for each wavelength.

For $\lambda = 5.890 \times 10^{-5} \text{ cm}$,

$$\begin{aligned} \text{radius of the ninth bright ring} &= \sqrt{\left(9 + \frac{1}{2}\right) \lambda R} \\ &= \sqrt{9.5 \times 5.890 \times 10^{-5} \times 100} \\ &= 0.236548 \text{ cm} \end{aligned}$$

$$\begin{aligned} \text{radius of the ninth dark ring} &= \sqrt{9 \lambda R} \\ &= 0.230239 \text{ cm} \end{aligned}$$

Similarly, for $\lambda = 5.896 \times 10^{-5} \text{ cm}$,

$$\begin{aligned} \text{radius of the ninth bright ring} &= \sqrt{9.5 \times 5.896 \times 10^{-5} \times 100} \\ &= 0.236669 \text{ cm} \end{aligned}$$

and

$$\begin{aligned} \text{radius of the ninth dark ring} &= \sqrt{9 \times 5.896 \times 10^{-5} \times 100} \\ &= 0.230356 \text{ cm} \end{aligned}$$

Thus the rings almost exactly superpose on each other. However, for large values of m , the two ring patterns may produce uniform illumination. To be more specific, when the air film thickness t is such that

$$2t = m\lambda_1 = \left(m + \frac{1}{2}\right)\lambda_2$$

$$\text{or} \quad \frac{2t}{\lambda_2} - \frac{2t}{\lambda_1} = \frac{1}{2} \quad (70)$$

then around that point the fringe system will completely disappear; i.e., the bright ring for the wavelength λ_1 will fall on the dark ring for the wavelength λ_2 and conversely. Thus the contrast will be zero and no fringe pattern will be visible. Rewriting Eq. (70) we get

$$2t \frac{\lambda_1 - \lambda_2}{\lambda_1 \lambda_2} = \frac{1}{2}$$

$$\begin{aligned} \text{or} \quad 2t &= \frac{1}{2} \frac{\lambda_1 \lambda_2}{\Delta \lambda} \approx \frac{1}{2} \frac{(5.893 \times 10^{-5})^2}{6 \times 10^{-8}} \\ &\approx 3 \times 10^{-2} \text{ cm} \end{aligned}$$

This will correspond to $m \approx 500$.

We shall see the effect of the same phenomenon if we slowly raise the convex lens in the upward direction as we had considered in Example 13.2. Let t_0 be the vertical distance through which the lens has been raised (see Fig. 13.27) and let t_0 be such that it satisfies the following equation:

$$\frac{2t_0}{\lambda_2} - \frac{2t_0}{\lambda_1} = \frac{1}{2}$$

or

$$t_0 = \frac{\lambda_1 \lambda_2}{4(\lambda_1 - \lambda_2)}$$

Thus, if the point J (see Fig. 13.27) corresponds to a dark spot for λ_1 then it will correspond to a bright spot for λ_2 and conversely. Further, the nearby dark rings for λ_1 , will almost fall at the same place as the bright rings for λ_2 and the interference pattern will be washed out. Thus viewing from a microscope we will not be able to see any ring pattern. Now, if the lens is further moved upwards by a distance t_0 , then we will have

$$\frac{2t_1}{\lambda_2} - \frac{2t_1}{\lambda_1} = 1 \quad (71)$$

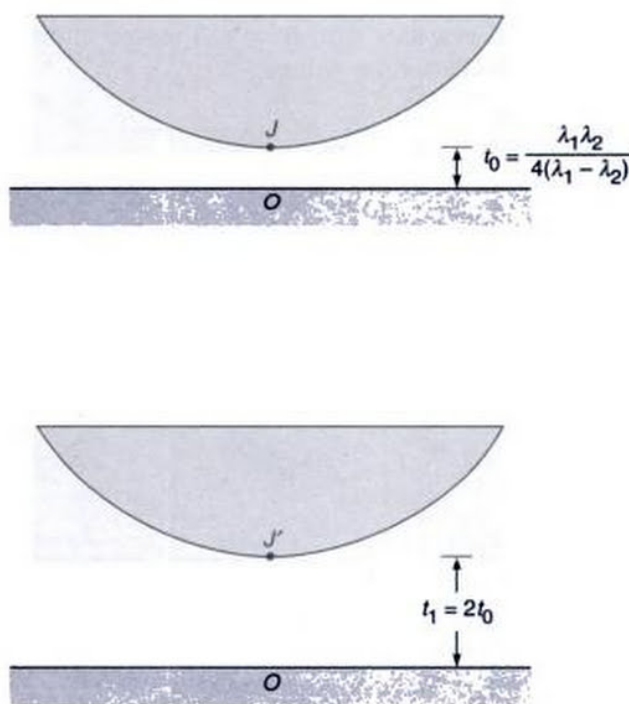


Fig. 13.27 In the Newton's rings experiment, if the light consists of two closely spaced wavelengths λ_1 and λ_2 (like the D_1 and D_2 lines of sodium) then, if the lens is separated by a distance $t_0 \left(= \frac{\lambda_1 \lambda_2}{4(\lambda_1 - \lambda_2)} \right)$ interference fringes will be washed out. The fringes will reappear when the distance is $2t_0$.

where $t_1 = 2t_0$. Consequently, if the point J' corresponds to a dark spot for λ_1 then it will also correspond to a dark spot for λ_2 . The fringe pattern will reappear but now with a slightly weaker contrast (see Sec. 15.8).

In this way if we continue to move the lens upwards the fringe system will reappear every time the lens is moved up by a distance $2t_0 \left(\approx \frac{1}{2} \frac{\lambda_1 \lambda_2}{\Delta \lambda} \right)$. This principle is used in a Michelson interferometer to measure the small wavelength difference $\Delta \lambda$, between two closely spaced lines (like the D_1 and D_2 lines of sodium).

It should be pointed out that for complete disappearance of the fringe pattern the intensities of the two lines λ_1 and λ_2 should be the same.

Another corollary of the above experiment consists in finding the change in the interference pattern (as we move up the convex lens) when we consider a single line of wavelength λ , but which has a width of $\Delta \lambda$. Thus we should assume all wavelengths between λ and $\lambda + \Delta \lambda$ to exist. By finding the approximate height at which the fringes disappear one can calculate $\Delta \lambda$. The coherence length (L) is related to $\Delta \lambda$ through the following relation (see Sec. 15.2):

$$L \sim \frac{\lambda^2}{\Delta \lambda} \quad (72)$$

13.11 THE MICHELSON INTERFEROMETER

A schematic diagram of the Michelson interferometer is shown in Fig. 13.28. S represents a light source (which may be a sodium lamp) and L represents a ground glass plate so that an extended source of almost uniform intensity is formed. G_1 is a beam splitter; i.e., a beam incident on G_1 gets partially reflected and partially transmitted. M_1 and M_2 are good quality plane mirrors having very high reflectivity. One of the mirrors (usually M_2) is fixed and the other (usually M_1) is capable of moving away or towards the glass plate G_1 along an accurately machined track by means of a screw. In the normal adjustment of the interferometer, the mirrors M_1 and M_2 are perpendicular to each other and G_1 is at 45° to the mirror.

Waves emanating from a point P get partially reflected and partially transmitted by the beam splitter G_1 , and the two resulting beams are made to interfere in the following manner: The reflected wave [shown as (1) in Fig. 13.28] undergoes a further reflection at M_1 and this reflected wave gets (partially) transmitted through G_1 ; this is shown as (5) in the figure. The transmitted wave [shown as (2) in Fig. 13.28] gets reflected by M_2 and gets (partially) reflected

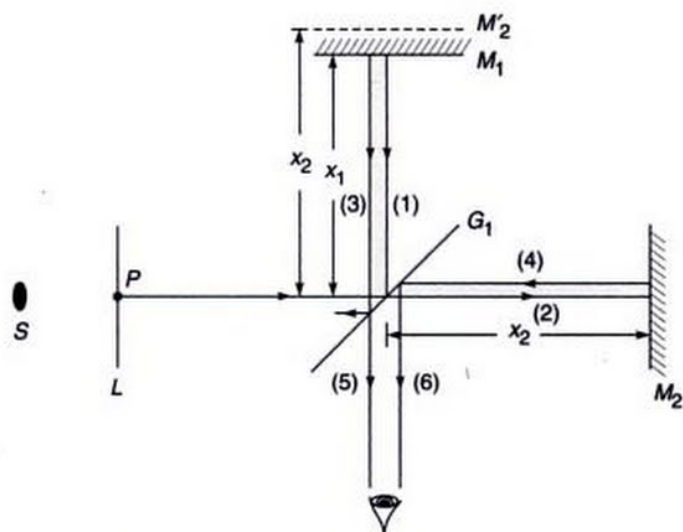


Fig. 13.28 Schematic of the Michelson interferometer.

by G_1 and results in the wave shown as (6) in the figure. Waves (5) and (6) interfere in a manner exactly similar to that shown in Fig. 13.16. This can easily be seen from the fact that if x_1 and x_2 are the distances of the mirrors M_1 and M_2 from the plate G_1 , then to the eye the waves emanating from the point P will appear to get reflected by two parallel mirrors [M_1 and M_2' – see Fig. 13.28] separated by a distance $(x_1 \sim x_2)$. As discussed in Sec. 13.7, if we use an extended source, then no definite interference pattern will be obtained on a photographic plate placed at the position of the eye. Instead, if we have a camera focused for infinity, then on the focal plane we will obtain circular fringes, each circle corresponding to a definite value of θ (see Figs 13.16 and 13.29); the circular fringes will look like the ones shown in Fig. 13.30. Now, if the beam splitter is just a simple glass plate, the beam reflected from the mirror M_2 will undergo an abrupt phase change of π (when getting reflected by the beam splitter) and since the extra path that one of the beams will traverse will be $2(x_1 \sim x_2)$, the condition for destructive interference will be

$$2d \cos \theta = m \lambda$$

where $m = 0, 1, 2, 3, \dots$ and

$$d = x_1 \sim x_2$$

and the angle θ represents the angle that the rays make with the axis (which is normal to the mirrors as shown in Fig. 13.29). Similarly, the condition for a bright ring would be

$$2d \cos \theta = \left(m + \frac{1}{2} \right) \lambda$$

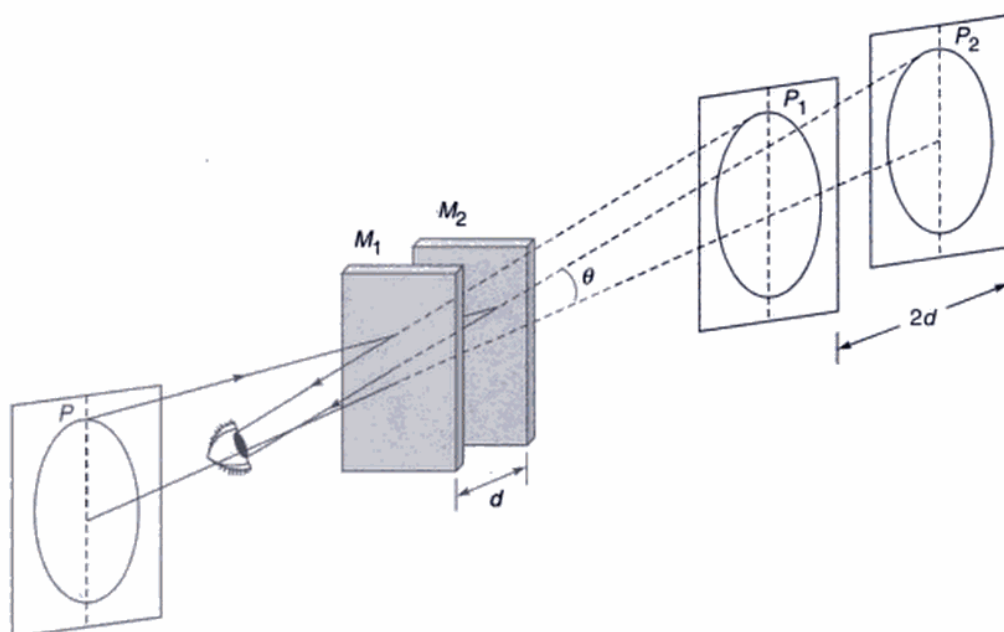


Fig. 13.29 A schematic of the formation of circular fringes [Adapted from Ref. 7].

For example, for $\lambda = 6 \times 10^{-5}$ cm if $d = 0.3$ mm, the angles at which the dark rings will occur will be

$$\theta = \cos^{-1} \left(\frac{m}{1000} \right)$$

$$= 0^\circ, 2.56^\circ, 3.62^\circ, 4.44^\circ, 5.13^\circ, 5.73^\circ, 6.28^\circ, \dots$$

corresponding to $m = 1000, 999, 998, 997, 996, 995, \dots$ Thus the central dark ring in Fig. 13.30(a) corresponds to $m = 1000$, the first dark ring corresponds to $m = 999$, etc. If we now reduce the separation between the two mirrors so that $d = 0.15$ mm, the angles at which the dark rings will occur will be [see Fig. 13.30(b)]

$$\theta = \cos^{-1} \left(\frac{m}{500} \right) = 0^\circ, 3.62^\circ, 5.13^\circ, 6.28^\circ, 7.25^\circ, \dots$$

where the angles now correspond to $m = 500, 499, 498, 497, 496, 495, \dots$ Thus as we start reducing the value of d , the fringes will appear to collapse at the centre and the fringes become less closely placed. It may be noted that if d is now slightly decreased, say from 0.15 mm to 0.14985 mm,

$$2d = 499.5 \lambda$$

the dark central spot in Fig. 13.30(b) (corresponding to $m = 500$) would disappear and the central fringe will become bright. Thus, as d decreases, the fringe pattern tends to collapse towards the centre. (Conversely, if d is increased, the fringe pattern will expand.) Indeed, if N fringes collapse to the centre as the mirror M_1 moves by a distance d_0 , then we must have

$$2d = m \lambda$$

$$2(d - d_0) = (m - N) \lambda$$

where we have put $\theta' = 0$ because we are looking at the central fringe. Thus

$$\lambda = \frac{2d_0}{N} \quad (77)$$

This provides us with a method for the measurement of the wavelength. For example, in a typical experiment, if one finds 1000 fringes collapse to the centre as the mirror is moved through a distance of 2.90×10^{-2} cm, then

$$\lambda = 5800 \text{ \AA}$$

The above method was used by Michelson for the standardization of the meter. He had found that the red cadmium line ($\lambda = 6438.4696 \text{ \AA}$) is one of the ideal monochromatic sources and as such this wavelength was used as a reference for the standardization of the meter. In fact he defined the meter by the following relation:

$$1 \text{ meter} = 1553164.13 \text{ red cadmium wavelengths,}$$

the accuracy is almost one part in 10^9 .

In an actual Michelson interferometer, the beam splitter G_1 consists of a plate (which may be about 1/2 cm thick), the back surface of which is partially silvered and the reflections occur at the back surface as shown in Fig. 13.31. It is immediately obvious that the beam (5) traverses the glass plate thrice and in order to compensate for this additional path, one introduces a 'compensating plate' G_2 which is exactly of the same thickness as G_1 . The compensating plate is not really necessary for a monochromatic source

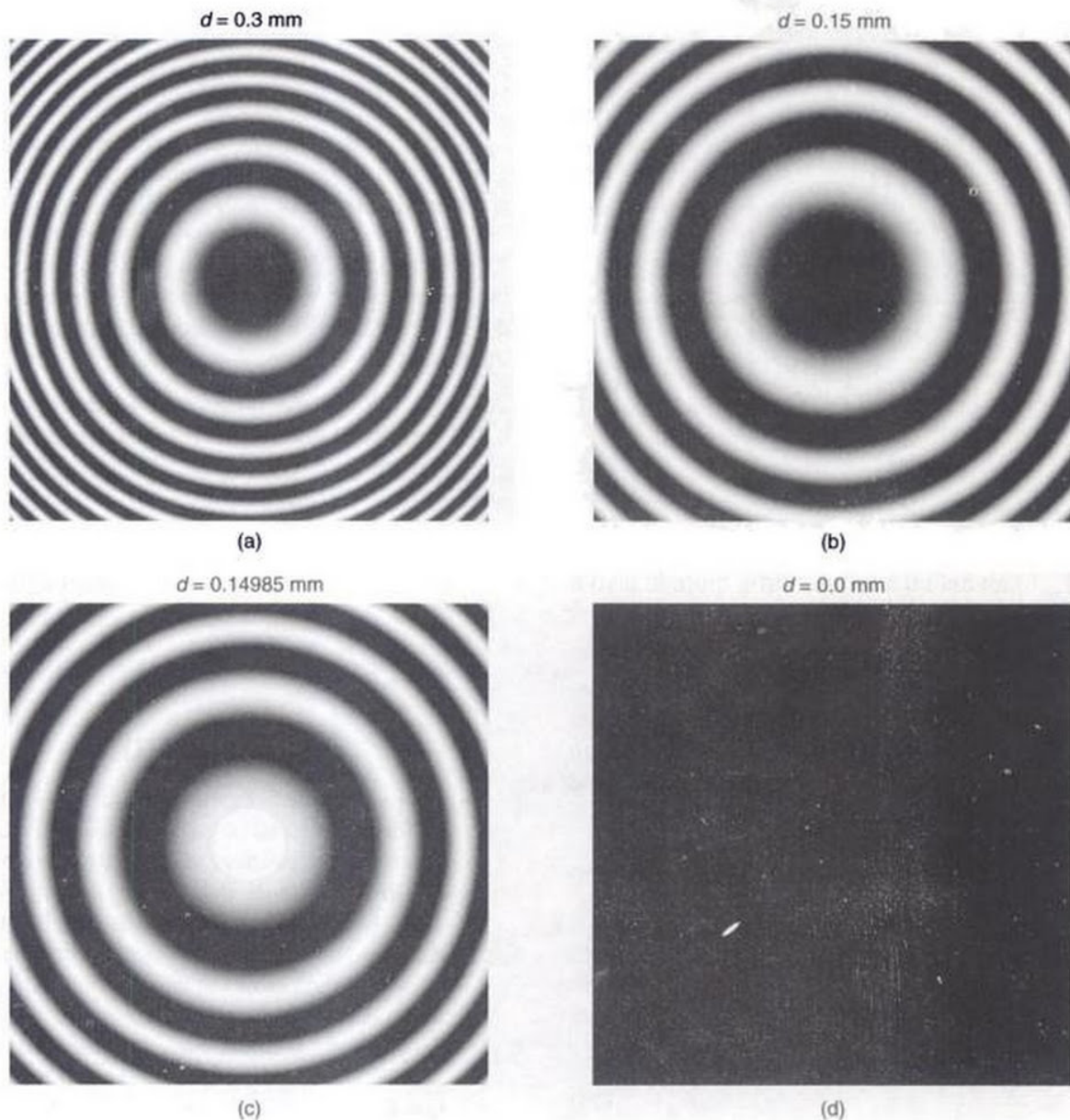


Fig. 13.30 Computer generated interference pattern produced by a Michelson interferometer.

because the additional path $2(n-1)t$ introduced by G_1 can be compensated by moving the mirror M_1 by a distance $(n-1)t$ where n is the refractive index of the material of the glass plate G_1 .

However, for a white light source it is not possible to simultaneously satisfy the zero path-difference condition for all wavelengths, since the refractive index depends on wavelength. For example, for $\lambda = 6560 \text{ \AA}$ and 4861 \AA , the refractive index of crown glass is 1.5244 and 1.5330 respectively. If we are using a 0.5 cm thick crown glass plate as G_1 , then M_1 should be moved by 0.2622 cm for $\lambda = 6560 \text{ \AA}$ and by 0.2665 cm for $\lambda = 4861 \text{ \AA}$, the difference between the two positions corresponding to over hundred

wavelengths! Thus, if we have a continuous range of wavelengths from 4861 \AA to 6560 \AA , the path difference between any pair of interfering rays (see Fig. 13.28) will vary so rapidly with wavelength that we would observe only a uniform white light illumination. However, in the presence of the compensating plate G_2 , one would observe a few coloured fringes around the point corresponding to zero path difference (see Sec. 12.9).

Michelson interferometer can also be used in the measurement of two closely spaced wavelengths. Let us assume that we have a sodium lamp which emits predominantly two closely spaced wavelengths 5890 \AA and 5896 \AA . The interferometer is first set corresponding to the zero path

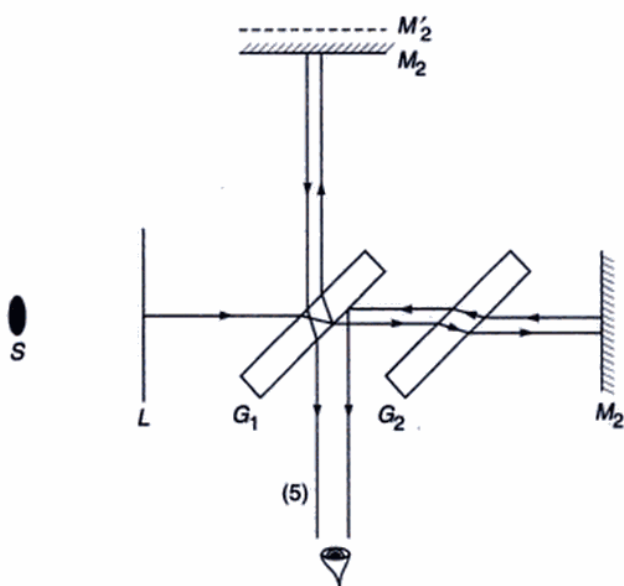


Fig. 13.31 In an actual interferometer there is also a compensating plate G_2 .

difference.* Near $d = 0$, both the fringe patterns will overlap. If the mirror M_1 is moved away (or towards) the plate G_1 through a distance d , then the maxima corresponding to the wavelength λ_1 will not, in general, occur at the same angle as λ_2 . Indeed, if the distance d is such that

$$\frac{2d}{\lambda_1} - \frac{2d}{\lambda_2} = \frac{1}{2} \quad (78)$$

and if $2d \cos \theta' = m\lambda_1$, then $2d \cos \theta' = (m + \frac{1}{2})\lambda_2$. Thus, the maxima of λ_1 will fall on the minima of λ_2 and conversely, and the fringe system will disappear. It can easily be seen that if

$$\frac{2d}{\lambda_1} - \frac{2d}{\lambda_2} = 1 \quad (79)$$

then interference pattern will again reappear. In general, if

$$\frac{2d}{\lambda_1} - \frac{2d}{\lambda_2}$$

is $1/2, 3/2, 5/2, \dots$ we will have disappearance of the fringe pattern and if it is equal to $1, 2, 3, \dots$ then the interference pattern will appear.

Instead of two discrete wavelengths, if the source consists of all wavelengths, lying between λ and $\lambda + \Delta\lambda$, then no interference pattern will be observed if

$$\frac{2d}{\lambda} - \frac{2d}{\lambda + \frac{\Delta\lambda}{2}} \geq \frac{1}{2}$$

or

$$2d \geq \frac{\lambda^2}{\Delta\lambda} \quad (80)$$

In this case the fringes will not reappear because we have a continuous range of wavelengths rather than two discrete wavelengths (see Sec. 15.2).

Example 13.4 For a sodium lamp, the distance traversed by the mirror between two successive disappearances is 0.289 mm. Calculate the difference in the wavelengths of the D_1 and the D_2 lines. Assume $\lambda = 5890 \text{ \AA}$.

Solution: When the mirror moves through a distance 0.289 mm, the additional path introduced is 0.578 mm. Thus

$$\frac{0.578}{\lambda} - \frac{0.578}{\lambda + \Delta\lambda} = 1$$

or

$$\Delta\lambda \approx \frac{\lambda^2}{0.578} = \frac{(5890 \times 10^{-7})^2}{0.578} \text{ mm} \\ \approx 6 \text{ \AA}$$

SUMMARY

- If a plane wave is incident normally on a thin film of uniform thickness d then the waves reflected from the upper surface interfere with the waves reflected from the lower surface. Indeed, for a film of thickness $\lambda/4n_f$ [where λ is the free space wavelength and n_f is the film refractive index which lies between the refractive indices of the two surrounding media], the wave reflected from the upper surface interferes destructively with the wave reflected from the lower surface and therefore the film acts like an antireflection layer.
- A medium consisting of a large number of alternate layers of high and low refractive indices of $n_0 + \Delta n$ and $n_0 - \Delta n$ of equal thickness d is called a periodic medium and the spatial period of the refractive index variation is denoted by $\Lambda (= 2d)$. For $\Delta n \ll n_0$, if $d \approx \frac{\lambda}{4n_0}$ (where λ is the free space wavelength), the reflections arising out of the individual reflections from the various interfaces would all be in phase and would result in a strong reflection. Thus for strong reflection at a chosen (free space) wavelength λ_B , the period of the refractive index variation should be

$$\Lambda = 2d = \frac{\lambda_B}{2n_0}$$

*The zero path difference is easily obtained by using white light where only a few coloured fringes, around $d = 0$, will be visible.

This is referred to as the Bragg condition. This is the principle of operation of Fiber Bragg gratings.

- If we place a plano-convex lens on a plane glass surface, a thin film of air is formed between the curved surface of the lens and the plane glass plate. If we allow monochromatic light (such as from a sodium lamp) to fall (almost normally) on the surface of the lens, then the light reflected from the curved surface interferes with the light reflected from the plane surface. Since the convex side of the lens is a spherical surface, the thickness of the air film will be constant over a circle and we will see concentric dark and bright rings. These rings are known as Newton's rings. The radii of the concentric rings are such that the difference between the square of the radii of successive fringes is very nearly a constant.
- The Michelson interferometer was used by Michelson for the standardization of the meter. He had found that the red cadmium line ($\lambda = 6438.4696 \text{ \AA}$) is one of the ideal monochromatic sources and as such this wavelength was used as a reference for the standardization of the meter. In fact he defined the meter by the following relation:
1 meter = 1553164.13 red cadmium wavelengths, the accuracy is almost one part in 10^9 .
- Michelson interferometer can also be used in the measurement of two closely spaced wavelengths.

PROBLEMS

- 13.1** A glass plate of refractive index 1.6 is in contact with another glass plate of refractive index 1.8 along a line such that a wedge of 0.5° is formed. Light of wavelength 5000 \AA is incident vertically on the wedge and the film is viewed from the top. Calculate the fringe spacing. The whole apparatus is immersed in an oil of refractive index 1.7. What will be the qualitative difference in the fringe pattern and what will be the new fringe width?
- 13.2** Two plane glass plates are placed on top of one another and on one side a cardboard is introduced to form a thin wedge of air. Assuming that a beam of wavelength 6000 \AA is incident normally, and that there are 100 interference fringes per centimetre, calculate the wedge angle.
- 13.3** Consider a non-reflecting film of refractive index 1.38. Assume that its thickness is $9 \times 10^{-6} \text{ cm}$. Calculate the wavelengths (in the visible region) for which the film will be non-reflecting. Repeat the calculations for the thickness of the film to be $45 \times 10^{-6} \text{ cm}$. Show that both the films will be non-reflecting for a particular wavelength but only the former one will be suitable. Why?
- 13.4** In the Newton's rings arrangement, the radius of curvature of the curved side of the plano-convex lens is 100 cm. For $\lambda = 6 \times 10^{-5} \text{ cm}$ what will be the radii of the 9th and 10th bright rings?
- 13.5** In the Newton's rings arrangement, the radius of curvature of the curved surface is 50 cm. The radii of the 9th and 16th dark rings are 0.18 cm and 0.2235 cm. Calculate the wavelength.
(Hint: The use of Eq. (66) will give wrong results, why?) (Ans. 5015 \AA)
- 13.6** In the Newton's rings arrangement if the incident light consists of two wavelengths 4000 \AA and 4002 \AA calculate the distance (from the point of contact) at which the rings will disappear. Assume that the radius of curvature of the curved surface is 400 cm.
(Ans. 4 cm)
- 13.7** In Problem 13.6 if the lens is slowly moved upward, calculate the height of the lens at which the fringe system (around the center) will disappear.
(Ans. 0.2 mm)
- 13.8** An equiconvex lens is placed on another equiconvex lens. The radii of curvatures of the two surfaces of the upper lens are 50 cm and those of the lower lens are 100 cm. The waves reflected from the upper and lower surface of the air film (formed between the two lenses) interfere to produce Newton's rings. Calculate the radii of the dark rings. Assume $\lambda = 6000 \text{ \AA}$.
[Ans. $0.0447 \sqrt{m} \text{ cm}$]
- 13.9** In the Michelson interferometer arrangement, if one of the mirrors is moved by a distance 0.08 mm, 250 fringes cross the field of view. Calculate the wavelength.
[Ans. 6400 \AA]
- 13.10** The Michelson interferometer experiment is performed with a source which consists of two wavelengths 4882 \AA and 4886 \AA . Through what distance does the mirror have to be moved between two positions of the disappearance of the fringes?
[Ans. 0.298 mm]
- 13.11** In the Michelson interferometer experiment, calculate the various values of θ' (corresponding to bright rings) for $d = 5 \times 10^{-3} \text{ cm}$. Show that if d is decreased to $4.997 \times 10^{-3} \text{ cm}$ the fringe corresponding to $m = 200$ disappears. What will be the corresponding values of θ' ? Assume $\lambda = 5 \times 10^{-5} \text{ cm}$.

REFERENCES AND SUGGESTED READINGS

1. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
2. M. Cagnet, M. Francon and S. Mallick, *Atlas of Optical Phenomena*, Springer-Verlag, Berlin, 1971.
3. E. F. Cave and L. V. Holroyd, 'Inexpensive Michelson interferometer', *Amer. J. Phys.*, Vol **23**, 61, 1955.
4. A. H. Cook, *Interference of Electromagnetic Waves*, Clarendon Press, Oxford, 1971.
5. M. Francon, *Optical Interferometry*, Academic Press, New York 1966.
6. A. K. Ghatak and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, London, 1989. [Reprinted by Foundation Books, New Delhi.]
7. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill Book Co., New York, 1976.
8. V. Oppenheim and J. H. Jaffe, 'Interference in an optical wedge' *Amer. J. Phys.*, Vol. **24**, 610, 1956.
9. J. Sladkova, *Interference of Light*, Iliffe Books Ltd., London, 1968.
10. W. H. Steel, *Interferometry*, Cambridge University Press, London, 1967.
11. S. Tolansky, *An Introduction to Interferometry*, Longmans Green and Co., London, 1955.

Chapter 14

Multiple Beam Interferometry

*When two Undulations ... coincide either perfectly or very nearly in Direction, their joint effect is a Combination of the Motions belonging to each.**

— Thomas Young (1801)

Important Milestone

1899 Marie Fabry and Jean Perot invented the Fabry–Perot interferometer which is characterized by a very high resolving power.

14.1 INTRODUCTION

In the last two chapters, we have been discussing interference between two beams which are derived from a single beam either by division of wavefront or by division of amplitude. In this chapter, we will discuss interference involving many beams which are derived from a single beam by multiple reflections (division of amplitude). Thus, for example, if a plane wave falls on a plane parallel glass plate, then the beam would undergo multiple reflections at the two surfaces and a large number of beams of successively diminishing amplitude will emerge on both sides of the plate. These beams (on either side) interfere to produce an interference pattern at infinity. We will show that the fringes so formed are much sharper than those by two beam interference and, therefore, the interferometers involving multiple beam interference have a high resolving power and hence find applications in high resolution spectroscopy.

14.2 MULTIPLE REFLECTIONS FROM A PLANE PARALLEL FILM

We consider the incidence of a plane wave on a plate of thickness h (and of refractive index n_2) surrounded by a medium of refractive index n_1 as shown in Fig. 14.1; as we will discuss later, the Fabry–Perot interferometer consists of two partially reflecting mirrors (separated by a fixed distance h) placed in air so that $n_1 = n_2 = 1$.

*The author found this quotation in Ref. 1.

Let A_0 be the (complex) amplitude of the incident wave. The wave will undergo multiple reflections at the two interfaces as shown in Fig. 14.1(a). Let r_1 and t_1 represent the amplitude reflection and transmission coefficients when the wave is incident from n_1 towards n_2 and let r_2 and t_2 represent the corresponding coefficients when the wave is incident from n_2 towards n_1 . Thus the amplitude of the successive reflected waves will be

$$A_0 r_1, A_0 t_1 r_2 t_2 e^{i\delta}, A_0 t_1 r_2^3 e^{2i\delta}, \dots$$

where
$$\delta = \frac{2\pi}{\lambda_0} \Delta = \frac{4\pi n_2 h \cos \theta_2}{\lambda_0} \quad (1)$$

represents the phase difference (between two successive waves emanating from the plate) due to the additional path traversed by the beam in the film (see Sec. 13.1) and in Eq. (1), θ_2 is the angle of refraction inside the film (of refractive index n_2), h the film thickness and λ_0 is the free space wavelength. Thus the resultant (complex) amplitude of the reflected wave will be

$$\begin{aligned} A_r &= A_0 [r_1 + t_1 t_2 r_2 e^{i\delta} (1 + r_2^2 e^{i\delta} + r_2^4 e^{2i\delta} + \dots)] \\ &= A_0 \left[r_1 + \frac{t_1 t_2 r_2 e^{i\delta}}{1 - r_2^2 e^{i\delta}} \right] \end{aligned} \quad (2)$$

Now, if the reflectors are lossless, the reflectivity and the transmittivity at each interface are given by (see Sec. 12.12)

$$\begin{aligned} R &= r_1^2 = r_2^2 \\ \tau &= t_1 t_2 = 1 - R \end{aligned}$$

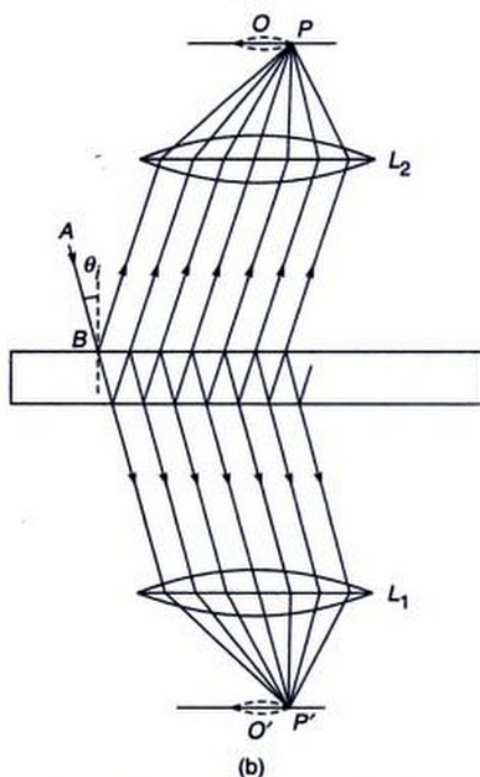
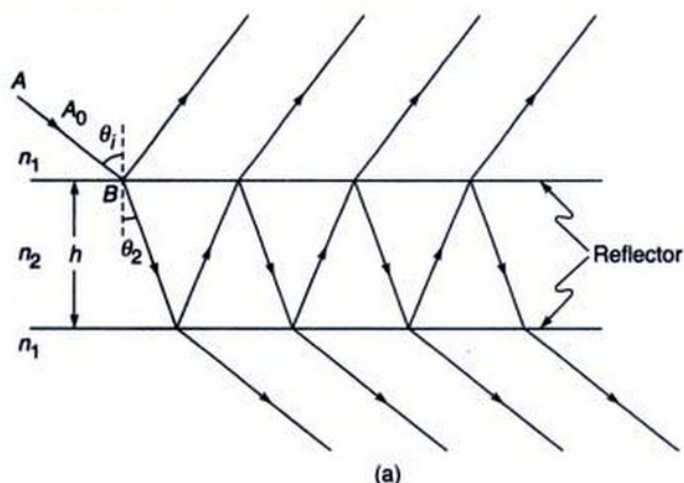


Fig. 14.1 (a) Reflection and transmission of a beam of amplitude A_0 incident at an angle θ_i on a film of refractive index n_2 and thickness h . (b) Any ray parallel to AB will focus at the same point P . If the ray AB is rotated about the normal at B , then the point P will rotate on the circumference of a circle centred at the point O ; this circle will be bright or dark depending on the value of θ_i . Rays incident at different angles will focus at different distances from the point O and one will obtain concentric bright and dark rings for an extended source.

(We are reserving the symbol T for the transmittivity of the Fabry-Perot etalon.) Thus

$$\frac{A_r}{A_0} = r_1 \left[1 - \frac{(1-R)e^{i\delta}}{1-Re^{i\delta}} \right]$$

where we have used the fact that $r_2 = -r_1$. Thus the reflectivity of the Fabry-Perot etalon is given by

$$\begin{aligned} \mathcal{R} &= \left| \frac{A_r}{A_0} \right|^2 = R \left| \frac{1-e^{i\delta}}{1-Re^{i\delta}} \right|^2 \\ &= R \frac{(1-\cos\delta)^2 + \sin^2\delta}{(1-R\cos\delta)^2 + R^2\sin^2\delta} \\ &= \frac{4R\sin^2\frac{\delta}{2}}{(1-R)^2 + 4R\sin^2\frac{\delta}{2}} \end{aligned}$$

or,

$$\mathcal{R} = \frac{F\sin^2\frac{\delta}{2}}{1+F\sin^2\frac{\delta}{2}} \quad (3)$$

where

$$F = \frac{4R}{(1-R)^2} \quad (4)$$

is called the coefficient of Finesse. One can immediately see that when $R \ll 1$, F is small and the reflectivity is proportional to $\sin^2\delta/2$. The same intensity distribution is obtained in the two beam interference pattern (see Sec. 12.7); we may mention here that we have obtained $\sin^2\delta/2$ instead of $\cos^2\delta/2$ because of the additional phase change of π in one of the reflected beams.

Similarly, the amplitude of the successive transmitted waves will be

$$A_0 t_1 t_2, A_0 t_1 t_2 r_2^2 e^{i\delta}, A_0 t_1 t_2 r_2^4 e^{2i\delta}, \dots$$

where, without any loss of generality, we have assumed the first transmitted wave to have zero phase. Thus the resultant amplitude of the transmitted wave will be given by

$$\begin{aligned} A_t &= A_0 t_1 t_2 [1 + r_2^2 e^{i\delta} + r_2^4 e^{2i\delta} + \dots] \\ &= A_0 \frac{t_1 t_2}{1 - r_2^2 e^{i\delta}} = A_0 \frac{1-R}{1-Re^{i\delta}} \end{aligned}$$

Thus the transmittivity T of the film is given by

$$T = \left| \frac{A_t}{A_0} \right|^2 = \frac{(1-R)^2}{(1-R\cos\delta)^2 + R^2\sin^2\delta}$$

or

$$T = \frac{1}{1 + F \sin^2 \frac{\delta}{2}} \quad (5)$$

It is immediately seen that the reflectivity and the transmittivity of the Fabry-Perot etalon add up to unity. Further,

$$T = 1$$

when

$$\delta = 2m\pi \quad ; \quad m = 1, 2, 3, \dots \quad (6)$$

In Fig. 14.2 we have plotted the transmittivity as a function of δ for different values of F . In order to get an estimate of the width of the transmission resources, let

$$T = \frac{1}{2} \quad \text{for} \quad \delta = 2m\pi \pm \frac{\Delta\delta}{2}$$

Thus

$$F \sin^2 \frac{\Delta\delta}{4} = 1 \quad (7)$$

The quantity $\Delta\delta$ represents the FWHM (Full Width at Half Maximum). In almost all cases, $\Delta\delta \ll 1$ and therefore, to a very good approximation, it is given by

$$\Delta\delta \approx \frac{4}{\sqrt{F}} = \frac{2(1-R)}{\sqrt{R}} \quad (8)$$

Thus the transmission resources become sharper as the value of F increases (see Fig. 14.2).

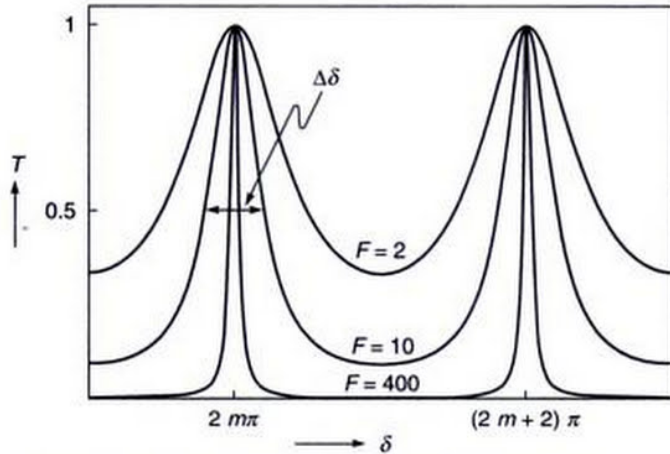


Fig. 14.2 The transmittivity of a Fabry-Perot etalon as a function of δ for different values of F ; the value of m is usually large. The transmission resonances become sharper as we increase the value of F . The FWHM (Full Width at Half Maximum) is denoted by $\Delta\delta$.

14.3 THE FABRY-PEROT ETALON

In this section, we will discuss the Fabry-Perot interferometer which is based on the principle of multiple beam interferometry discussed in the last section. The interferometer (as shown in Fig. 14.3) consists of two plane glass (or quartz) plates which are coated on one side with a partially reflecting metallic film* (of aluminum or silver) of about 80% reflectivity. These two plates are kept in such a way that they enclose a plane parallel slab of air between their coated surfaces. If the reflecting glass plates are held parallel to each other at a fixed separation, we have what is known as a Fabry-Perot etalon. In fact, we may neglect the presence of the plates and consider only the reflection (and transmission) by the metallic film; further, if the plates are parallel, the rays will not undergo any deviation.

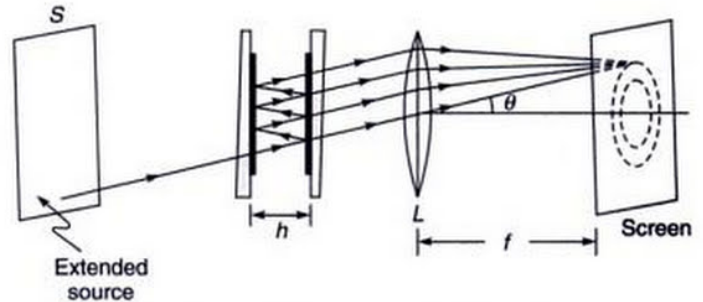


Fig. 14.3 The Fabry-Perot etalon

In a typical experiment, light from a broad source is collimated by a lens and is passed through the Fabry-Perot etalon as shown in Fig. 14.3. Thus, if we consider light of a specific wavelength λ_0 , the incident light will be completely transmitted (i.e., $T = 1$) if the angle of incidence is such that

$$\delta = \frac{4\pi}{\lambda_0} n_2 h \cos \theta_2 = 2m\pi \quad (9)$$

or

$$\cos \theta_2 = \frac{m\lambda_0}{2n_2 h} \quad (10)$$

For large values of F , when θ_2 is slightly different from the value given by the above equation, the transmittivity will be very small. Hence, for a given wavelength, at the focal plane of the lens L , we will obtain a fringe pattern consisting of concentric rings — each bright ring will correspond to a particular value of m . The sharpness of the bright rings (and hence the resolving power of the etalon) will increase with the value of F .

*In the visible region of the spectrum, silver is the best metal to coat with (the reflectivity is about 0.97 in the red region and decrease to about 0.90 in the blue region). But beyond the blue region, the reflectivity falls rapidly. Aluminum is usually employed below 4000 Å.

Example 14.1 As an example, we assume an etalon with $n_2 = 1$, $h = 1$ cm and $F = 400$ ($F = 400$ implies $R \approx 0.905$; i.e., each mirror of the etalon has about 90% reflectivity). In Fig. 14.4 we have plotted the intensity variation with θ for $\lambda_0 = 5000 \text{ \AA}$ and 4999.98 \AA . The actual fringe pattern (as obtained on the focal plane of a lens of focal length 25 cm) is shown in Fig. 14.5. Now, for

$$\lambda_0 = \lambda_1 = 5000 \text{ \AA}$$

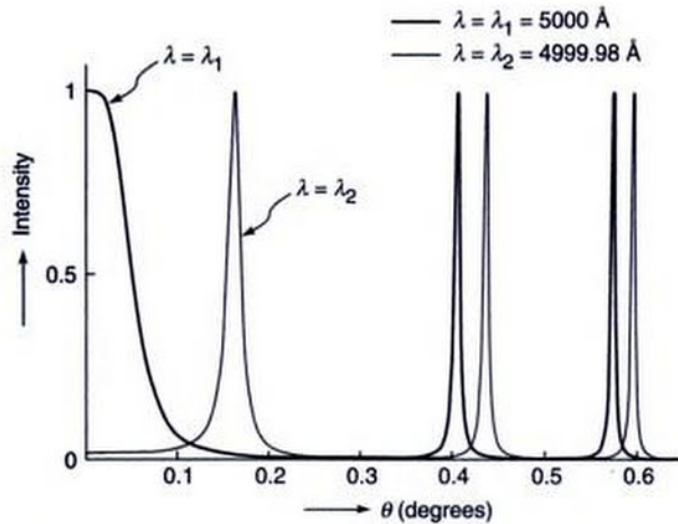


Fig. 14.4 The variation of intensity with θ for a Fabry-Perot interferometer with $n_2 = 1$, $h = 1.0$ cm and $F = 400$, corresponding to $\lambda_0 = 5000 \text{ \AA}$ ($=\lambda_1$) and $\lambda_0 = 4999.98 \text{ \AA}$ ($=\lambda_2$).



Fig. 14.5 The (computer generated) ring pattern as obtained (on the focal plane of a lens) in a Fabry-Perot etalon with $n_2 = 1$, $h = 1.0$ cm and $F = 400$, corresponding to $\lambda_0 = 5000 \text{ \AA}$ ($=\lambda_1$) and $\lambda_0 = 4999.98 \text{ \AA}$ ($=\lambda_2$).

Eq. (9) gives us

$$\theta_2 = \cos^{-1} \left(\frac{m}{40000} \right)$$

Thus bright rings will form at

$$\theta_2 = 0^\circ, 0.41^\circ, 0.57^\circ, 0.70^\circ, \dots$$

corresponding to $m = 40000, 39999, 39998, 39997, \dots$ respectively. This is shown as the thick curve in Fig. 14.4. On the other hand, for

$$\lambda_0 = \lambda_2 = 4999.98 \text{ \AA}$$

we get

$$\theta_2 = \cos^{-1} \left(\frac{m}{40000.16} \right)$$

Thus bright rings will form at

$$\theta_2 = 0.162^\circ, 0.436^\circ, 0.595^\circ, \dots$$

corresponding to $m = 40000, 39999$ and 39998 respectively. This is shown as the thin curve in Fig. 14.4. The corresponding ring patterns as obtained on the focal plane of the lens is shown in Fig. 14.5; from the figure we can see that the two spectral lines having a small wavelength difference of 0.02 \AA are quite well resolved by the etalon. In the figure, the central bright spot and the first ring corresponds respectively to $\lambda_0 = 5000 \text{ \AA}$ and $\lambda_0 = 4999.98 \text{ \AA}$; both corresponding to $m = 40000$. The next two closely spaced rings correspond to $m = 39999$ for the two wavelengths.

14.3.1 Flatness of the Coated Surfaces

In order to have sharp fringes, the coated surfaces should be parallel to a very high degree of accuracy. Indeed, the coated surfaces should be flat within about $\lambda/50$ where λ is the wavelength of light. In order to see this, we assume that in the above example h is increased by $\lambda/20$ ($=250 \text{ \AA} = 2.5 \times 10^{-6} \text{ cm}$):

$$h = 1 + 2.5 \times 10^{-6} = 1.0000025 \text{ cm}$$

For $\lambda_0 = 5000 \text{ \AA}$, we will have

$$\theta_2 = \cos^{-1} \left(\frac{m}{40000.1} \right)$$

and bright rings will form at

$$\theta_2 = 0.128^\circ, 0.425^\circ, 0.587^\circ, \dots$$

If we compare the results obtained in Example 14.1, we will find that if there is a variation in the spacing by about $\lambda/20$, the fringes corresponding to the wavelengths 5000 \AA and 4999.98 \AA will start overlapping. Thus the coated surfaces should be parallel within a very small fraction of

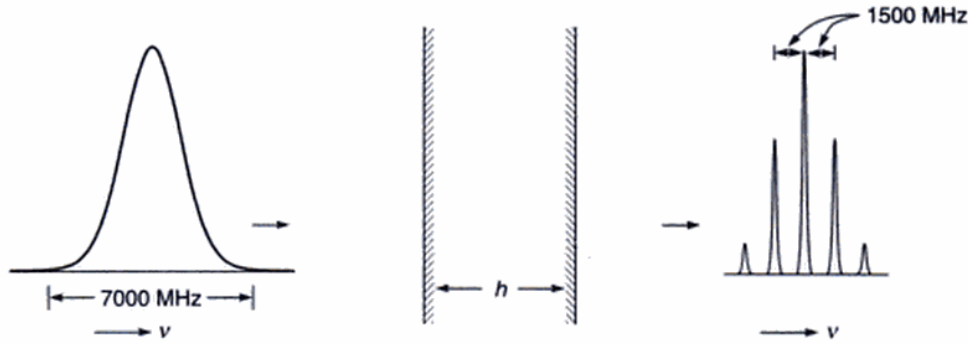


Fig. 14.6 A beam having a spectral width of about 7000 MHz (around $\nu_0 = 6 \times 10^{14}$ Hz) is incident normally on a Fabry-Perot etalon with $h = 10$ cm and $n_2 = 1$. The output has five narrow spectral lines.

the wavelength. Further, the two uncoated surfaces of each plate are made to have a slight angle between them (~ 1 to 10 minutes – see Fig. 14.3) so that one could avoid the unwanted fringes formed due to multiple reflections in the plate itself.

14.3.2 Modes of the Fabry-Perot Cavity

We consider a polychromatic beam incident normally ($\theta_2 = 0$) on a Fabry-Perot etalon with air between the reflecting plates ($n_2 = 1$) – see Fig. 14.6. In terms of the frequency

$$\nu = \frac{c}{\lambda_0}$$

Eq. (9) tells us that transmission resonance will occur when

$$\nu = \nu_m = m \frac{c}{2h} \quad (11)$$

where m is an integer. The above equation represents the different (longitudinal) modes of the (Fabry-Perot) cavity. For $h = 10$ cm, the frequency spacing of two adjacent modes would be given by

$$\delta\nu = \frac{c}{2h} = 1500 \text{ MHz}$$

For an incident beam having a central frequency of

$$\nu = \nu_0 = 6 \times 10^{14} \text{ Hz}$$

and a spectral width* of 7000 MHz the output beam will have frequencies

$$\nu_0, \nu_0 \pm \delta\nu \text{ and } \nu_0 \pm 2 \delta\nu$$

as shown in Fig. 14.6. One can readily calculate from Eq. (11) that the five lines correspond to

$$m = 399998, 399999, 400000, 400001 \text{ and } 400002$$

Figure 14.7 shows a typical output of a multilongitudinal (MLM) laser diode. The wavelength spacing between two modes is about $0.005 \mu\text{m}$.

14.4 THE FABRY-PEROT INTERFEROMETER

If one of the mirrors is kept fixed while the other is capable of moving to change the separation between the two

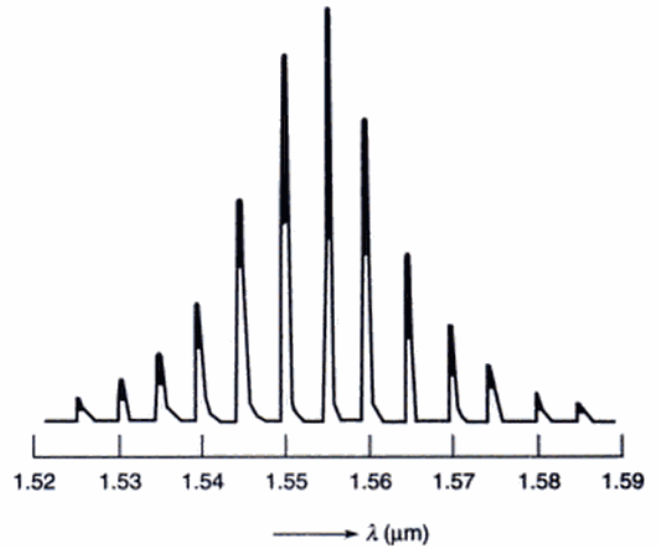


Fig. 14.7 Typical output spectrum of a Fabry-Perot multilongitudinal mode (MLM) laser diode; the wavelength spacing between two modes is about $0.005 \mu\text{m}$ (After Ref. 9).

*For $\nu_0 = 6 \times 10^{14}$ Hz, $\lambda_0 = 5000 \text{ \AA}$ and a spectral width of 7,000 MHz would imply $\left| \frac{\Delta\lambda_0}{\lambda_0} \right| = \left| \frac{\Delta\nu}{\nu_0} \right| = \frac{7 \times 10^9}{6 \times 10^{14}} = 1.2 \times 10^{-5}$ giving $\Delta\lambda_0 = 0.06 \text{ \AA}$. Thus a frequency spectral width of 7000 MHz (around $\nu_0 = 6 \times 10^{14}$ Hz) implies a wavelength spread of only 0.06 \AA .

mirrors, the system is called a Fabry–Perot interferometer. For a beam incident normally on the interferometer, we vary the separation h and measure the intensity variation on the focal plane of the lens L as shown in Fig. 14.8. Such an arrangement is usually referred to as a scanning Fabry–Perot interferometer. Since the separation h is varied, we write it as

$$h = h_0 + x \quad (12)$$

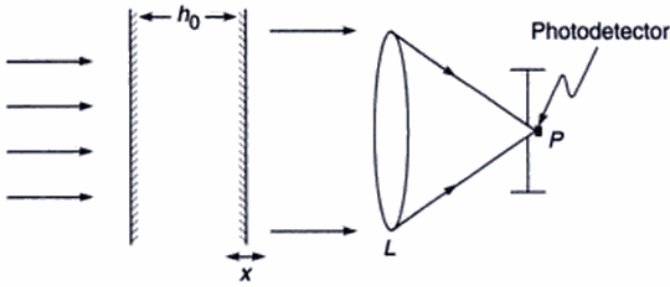


Fig. 14.8 A scanning Fabry–Perot interferometer. The intensity variation is recorded (by a photodetector) on the focal plane of the lens L .

If the incident beam is monochromatic, a typical variation of intensity at the point P is shown in Fig. 14.9. The figure corresponds to the frequency of the incident beam being

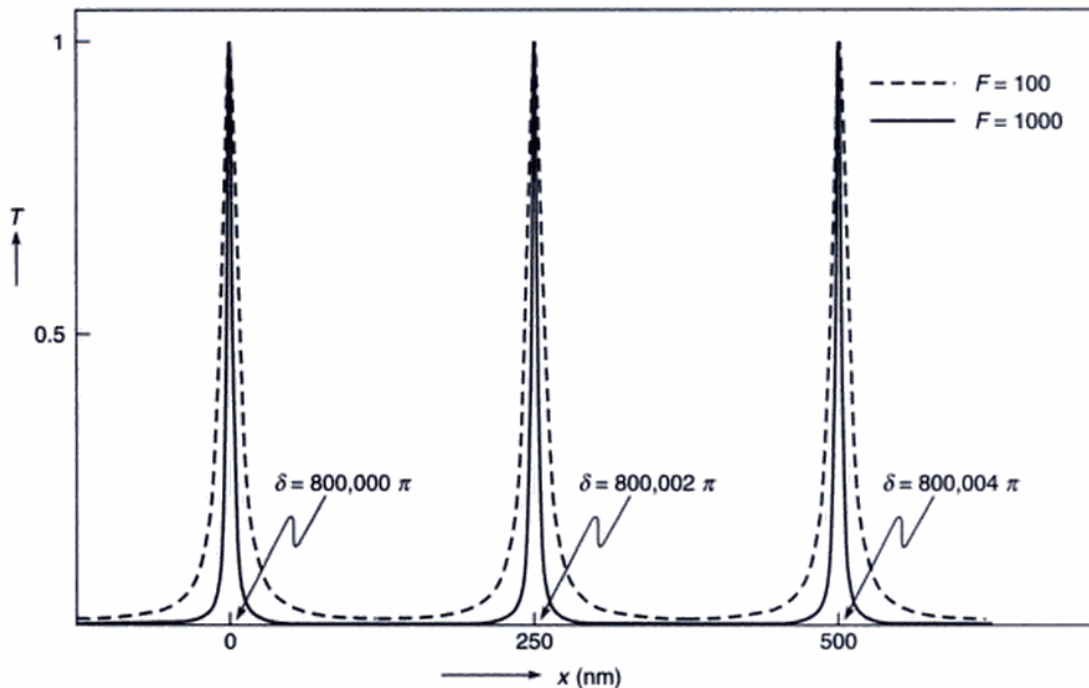


Fig. 14.9 Variation of intensity at the point P with x (see Fig. 14.8) for a monochromatic beam incident normally on a scanning Fabry–Perot interferometer; the solid curve corresponds to $F = 1000$ and the dashed curve corresponds to $F = 100$.

$$\nu = \nu_0 = 6 \times 10^{14} \text{ Hz}$$

For $h_0 = 10 \text{ cm}$, $n_2 = 1$ and $\cos \theta_2 = 1$, we get

$$\begin{aligned} \delta &= \frac{4\pi \nu_0 (h_0 + x)}{c} \\ &= 800000 \pi \left(1 + \frac{x}{h_0} \right) \end{aligned}$$

Thus transmittivity resonances will occur for

$$\delta = 800000\pi, 800002\pi, 800004\pi, \dots$$

which will occur when

$$x = 0, 250 \text{ nm}, 500 \text{ nm}, \dots$$

respectively. The two curves in Fig. 14.9 correspond to $F = 100$ and $F = 1000$. Notice that the transmission resonances become sharper if we increase the value of F . Figure 14.10 shows variation of intensity at the point P when the incident beam has two frequencies separated by 300 MHz. Obviously, the two frequencies are well resolved.

We may mention here that if the frequency of the incident beam is increased by $c/2h_0$, i.e., if

$$\nu = \nu_0 + \frac{c}{2h_0}$$

then one can easily show that transmission resonances will occur at the same values of x , the corresponding values of

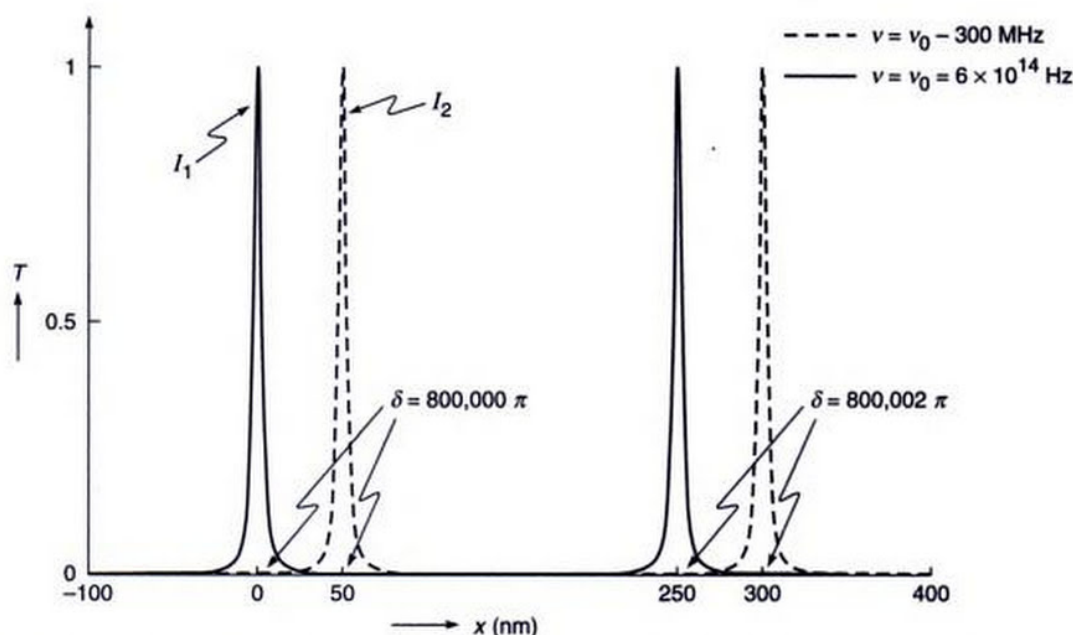


Fig. 14.10 Variation of intensity at the point P with x (see Fig. 14.8) when the incident beam has two frequencies separated by 300 MHz.

δ will be 800002π (corresponding to $x = 0$), 800004π (corresponding to $x = 250$ nm) etc. Indeed if

$$\nu = \nu_0 \pm p \frac{c}{2h_0}; \quad p = 1, 2, 3, \dots$$

we will have the same T vs. x curve. The quantity

$$\Delta\nu_s = \frac{c}{2h_0} \quad (13)$$

is known as the free spectral range (FSR) of the interferometer. Thus when the spectrum has widely separated wavelength components, we will have overlapping of orders.

14.5 RESOLVING POWER

We will first consider the resolving power corresponding to a beam incident normally on a scanning Fabry-Perot interferometer. This will be followed by the case corresponding to the Fabry-Perot etalon.

14.5.1 Resolving Power of a Scanning Fabry-Perot Interferometer

We consider the presence of two frequencies ν_1 and ν_2 of equal intensity in the beam incident normally on a scanning Fabry-Perot interferometer. For the two frequencies to be just resolved, we assume that the half intensity point of ν_1 falls on the half intensity point of ν_2 as shown in Fig. 14.11.

When this happens, the minimum of the resultant intensity distribution (shown as the dashed curve in Fig. 14.11) is about 74% of the corresponding maximum value. Now, as discussed in Sec. 14.2, if the half intensity point occurs at

$$\delta = \delta_{1/2} = 2m\pi \pm \frac{\Delta\delta}{2} \quad (14)$$

then

$$\Delta\delta = \frac{4}{\sqrt{F}} \quad (15)$$

[see Eq. (8)]. Consider the frequency ν_1 . If the intensity maximum occurs at $h = h_1$ then

$$\delta_1 = \frac{4\pi h_1 \nu_1}{c} = 2m\pi \quad (16)$$

Let the intensity maximum for $\nu = \nu_2 (= \nu_1 + \Delta\nu_1)$ occurs at

$$h = h_2 = h_1 + \Delta h_1$$

Thus

$$\delta_2 = \frac{4\pi (h_1 + \Delta h_1) (\nu_1 + \Delta\nu_1)}{c} = 2m\pi \quad (17)$$

Using Eqs (16) and (17) and neglecting the second order term $\Delta h_1 \Delta\nu_1$, we get

$$\nu_1 \Delta h_1 + h_1 \Delta\nu_1 = 0$$

or

$$\Delta h_1 = -\frac{h_1}{\nu_1} \Delta\nu_1 \quad (18)$$

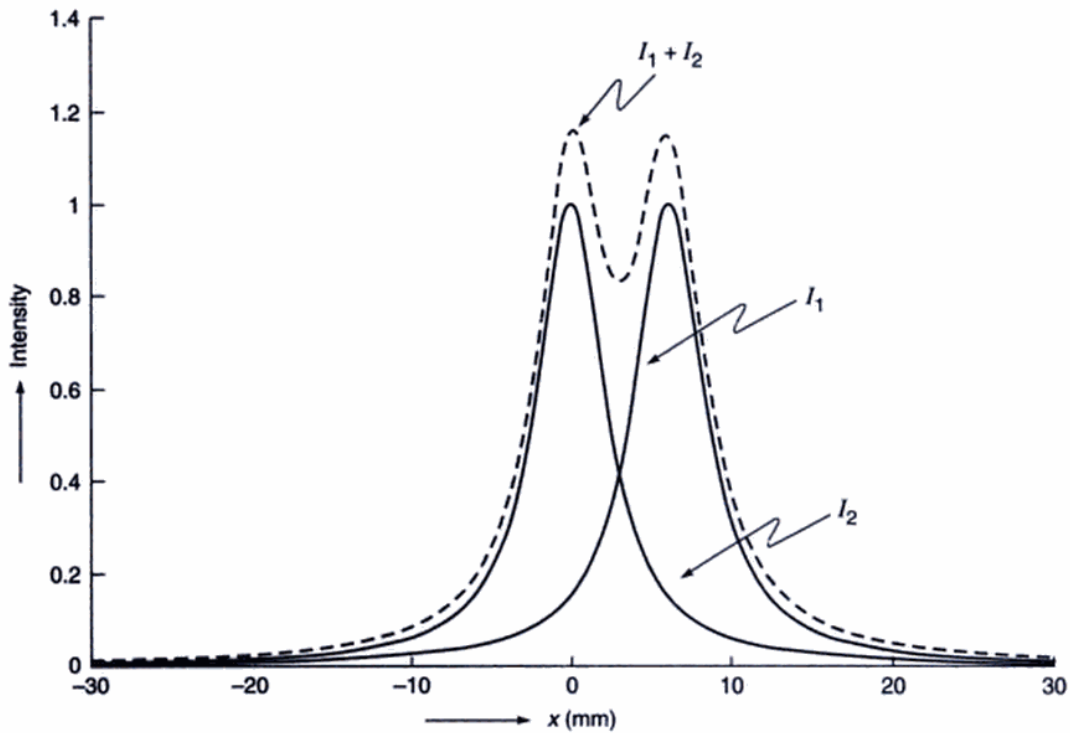


Fig. 14.11 The individual intensity variations I_1 and I_2 in the presence of two frequencies ν_1 and ν_2 and the total intensity variation ($I_1 + I_2$) when the two frequencies are just resolved.

Equation (18) implies that for Δh_1 to be positive, $\Delta \nu_1$ should be negative. Now, for the frequency ν_1 , let the half intensity point occur at $h = h_1 + \delta h_1$ (the corresponding value of δ will be $2m\pi + \frac{1}{2} \Delta \delta_1$; thus using Eq. (16)

$$\frac{4\pi \nu_1 \delta h_1}{c} = \frac{1}{2} \Delta \delta_1 \approx \frac{2}{\sqrt{F}} \quad (19)$$

or

$$\Delta h_1 \approx \frac{c}{2\pi \nu_1 \sqrt{F}} \quad (20)$$

For the two frequencies to be just resolved

$$\Delta h_1 = 2 \delta h_1 \approx \frac{c}{\pi \nu_1 \sqrt{F}} \quad (21)$$

Using Eq. (18) we get for the resolving power

$$\left| \frac{\nu_1}{\Delta \nu} \right| = \frac{h_1}{\Delta h} = \frac{\pi h_1 \nu_1 \sqrt{F}}{c}$$

Or, dropping the subscript we get

$$\text{Resolving power} = \left| \frac{\nu}{\Delta \nu} \right| = \frac{\pi h \nu \sqrt{F}}{c} \quad (22)$$

Or, in terms of the wavelength

$$\text{Resolving power} = \left| \frac{\lambda_0}{\Delta \lambda_0} \right| = \frac{\pi h \sqrt{F}}{\lambda_0} \quad (23)$$

For $h = 1$ cm, $\lambda_0 = 6 \times 10^{-5}$ cm

$$\Delta \lambda \approx 0.013 \text{ \AA for } F = 80$$

$$\approx 0.006 \text{ \AA for } F = 360$$

14.5.2 Resolving Power of a Fabry-Perot Etalon

We consider light from a broad source incident on a Fabry-Perot etalon as shown in Fig 14.3. We once again consider the presence of two wavelengths λ_1 and λ_2 of equal intensity. Now, $T = 1$ if the angle of incidence is such that [see Eq. (9)]

$$\delta = \frac{4\pi \nu}{c} h \mu = 2 m \pi \quad (24)$$

where $\mu = \cos \theta$, and for the sake of simplicity, we have dropped the subscript on μ and θ . We can now have arguments very similar to that in Sec. 14.5.1 except now h is fixed and μ ($= \cos \theta$) is varied. Thus, if the m^{th} order intensity maxima for $\nu = \nu_1$ and $\nu = \nu_2$ ($= \nu_1 + \Delta \nu_1$) occur at $\mu = \mu_1$ and $\mu = \mu_2$ ($= \mu_1 + \Delta \mu_1$), then

$$\delta_1 = \frac{4\pi \nu_1 h \mu_1}{c} = 2 m \pi \quad (25)$$

and

$$\delta_2 = \frac{4\pi h (\nu_1 + \Delta \nu_1) (\mu_1 + \Delta \mu_1)}{c} = 2 m \pi \quad (26)$$

Thus, neglecting the second order term we get

$$\Delta\mu_1 = -\frac{\mu_1}{v_1} \Delta v_1 \quad (27)$$

Now, for the frequency v_1 , let the half intensity point occur at $\mu = \mu_1 + \delta\mu_1$ (the corresponding value of δ will be $2m\pi + \frac{1}{2}\Delta\delta_1$); thus using Eq. (24)

$$\frac{4\pi v_1 h \delta\mu_1}{c} = \frac{1}{2} \Delta\delta_1 = \frac{2}{\sqrt{F}} \quad (28)$$

or

$$\delta\mu_1 \approx \frac{c}{2\pi v_1 h \sqrt{F}} \quad (29)$$

As discussed earlier, for the two frequencies to be just resolved, we assume that the half intensity point of v_1 falls on the half intensity point of v_2 giving

$$\Delta\mu_1 = 2\delta\mu_1 \approx \frac{c}{\pi v_1 h \sqrt{F}} \quad (30)$$

Using Eq. (27) we get

$$\text{Resolving power} = \left| \frac{v_1}{\Delta v_1} \right| = \frac{\mu_1}{\Delta\mu_1} = \frac{\pi v_1 h \sqrt{F} \mu_1}{c} \quad (31)$$

Or, in terms of the wavelength

$$\text{Resolving power} = \left| \frac{\lambda_0}{\Delta\lambda_0} \right| = \frac{\pi h \sqrt{F} \cos \theta}{\lambda_0} \quad (32)$$

Thus for $F = 360$ ($R = 0.9$), $h = 1$ cm, $\lambda_0 = 5000$ Å

$$\left| \frac{\lambda_0}{\Delta\lambda_0} \right| \approx 1.2 \times 10^6$$

where we have assumed normal incidence. The above equation gives

$$\Delta\lambda_0 \approx 0.004 \text{ Å}$$

Thus a Fabry-Perot instrument can resolve wavelengths differing by about 10^{-3} Å. This is in contrast to that of a grating (say having 25000 grooves) which resolves up to about 0.1 Å at $\lambda = 5000$ Å and that of a prism (made of dense flint glass with 5 cm base) which resolves only up to about 1 Å at 5000 Å. It must be noted that in the above analysis, we have considered two monochromatic lines at λ and $\lambda + \Delta\lambda$. In general, the lines at the two wavelengths λ and $\lambda + \Delta\lambda$ themselves will have a wavelength spread and this restricts the use of such high resolving powers.

When the Fabry-Perot interferometer is used to analyze spectra with closely spaced lines, then the distance between the adjacent maxima would be greater than the displacement between the system of rings of the spectral lines. But when the spectrum has widely separated wavelength components,

then it might happen that the displacement between the rings is greater than the separation between adjacent maxima. The results in the 'overlapping' of orders (see also the discussion at the end of Sec 14.4). The difference in wavelength ($\Delta\lambda_s$) which corresponds to a displacement of one order, is called the spectral range of the interferometer. Thus we can write

$$\Delta\lambda_s = \frac{\lambda^2}{2nh \cos \theta} \quad (33)$$

This becomes, for near normal incidence ($\theta \approx 0$),

$$\Delta\lambda_s = \frac{\lambda^2}{2nh} \quad (34)$$

which is found to be inversely proportional to h . This is in contrast to the resolving power which depends directly on h [see Eqs (31) and (32)].

When the spectrum is complex consisting of a number of widely separated wavelength components, each with a hyperfine structure, then one can separate the different wavelength components by employing the Fabry-Perot interferometer along with a spectrograph as shown in Fig. 14.12(a). The light emerging from the source S is rendered parallel by the lens L_1 . The interference pattern formed by the Fabry-Perot interferometer (marked by FP in the figure) is made to fall on the slit of the spectrograph. The spectrograph separates the spectral components and one obtains in the plane P images of the slit, each crossed by fringes as shown in Fig. 14.12(b).

14.6 THE LUMMER-GEHRCKE PLATE*

We saw in Sec. 14.2 that the sharpness of fringes (and hence the resolving power) of a multiple beam interferometer increases as the reflectivity R of the plate increases. But one cannot use every thick coating of metals to increase the reflectivity as the intensity of the beam would be reduced considerably due to absorption in metallic coatings. This difficulty can be overcome by the use of the phenomenon of total internal reflection (instead of metallic reflection); this is used in the Lummer-Gehrcke plate which will be discussed in this section.

A Lummer-Gehrcke plate is a plane parallel made of glass (or quartz), on one end of which a small right-angled prism of the same material is fixed (see Fig. 14.13). The angle of the prism is chosen in such a way that the rays incident normally on the surface of the prism hit the two surfaces of the plate at an angle slightly less than the critical angle.** Since the two surfaces are parallel, all successive reflections will occur at the same (near critical) angle. Most

*Sections 14.6 and 14.7 have been very kindly written by Professor Anurag Sharma.

**Beyond the critical angle, the reflection is total while slightly below the critical angle, the reflectivity is high (see Sec. 21.2).

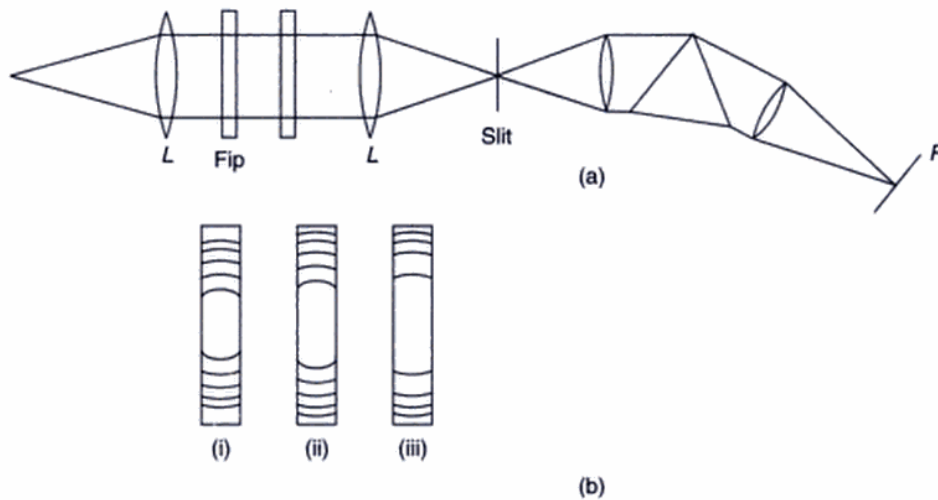


Fig. 14.12 (a) A Fabry-Perot interferometer used in conjunction with a spectrograph. (b) The interlaced fringes formed in the plane of the slit are separated by the prism. For example, (i), (ii) and (iii) may correspond to the lines in the red, yellow and green regions respectively as observed on plane P .

of the light will be reflected with a little fraction being transmitted at each reflection. Thus, there will emerge from the upper and lower surfaces of the plate a series of waves which would finally interfere to produce interference fringes in the plane P (see Fig. 14.13). Notice that the prism suppresses the externally reflected beam. In the plane P , one obtains fringe patterns on either side of the plate. The fringes are approximately straight lines parallel to the plate surfaces.

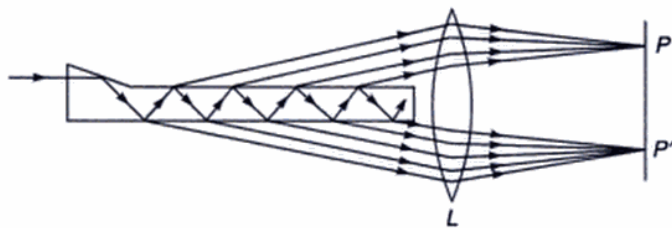


Fig. 14.13 The Lummer-Gehrcke plate.

We will not go into the details of the theory of the Lummer-Gehrcke plate but two points may be noted:

- (a) Unlike in the case of Fabry-Perot interferometer, the space between the reflecting surfaces is a dispersive medium, and
- (b) The number of reflections is also not very large as in the case of the Fabry-Perot interferometer; the number of reflections depends on the length of the plate and the angle θ , (see Fig. 14.13). Thus, the resolving power of the instrument depends on the length of the plate.

Earlier, Lummer-Gehrcke plates were used in high resolution spectroscopy. However, it has been replaced by the more flexible Fabry-Perot interferometer.

14.7 INTERFERENCE FILTERS

When a Fabry-Perot interferometer is illuminated by a monochromatic (uncollimated) beam, we get a spectrum consisting of different intensity maxima which satisfy the following relation:

$$2 n h \cos \theta_r = m \lambda \quad (35)$$

Now if a Fabry-Perot interferometer is illuminated with a collimated white light incident normally ($\theta_r = 0$), maxima of different orders are formed in the transmitted light corresponding to wavelengths given by

$$\lambda = \frac{2 n h}{m} \quad (36)$$

If h is large, a large number of maxima will be observed in the visible region; for example, about 23,000 maxima are observed if $h = 1$ cm. But, if we go on reducing h , we reach a situation in which only one or two maxima are obtained in the visible region. For example, if $n = 1.5$ and $h = 6 \times 10^{-5}$ cm, there are only two maxima in the visible region, corresponding to $\lambda = 6000 \text{ \AA}$ ($m = 3$) and $\lambda = 4500 \text{ \AA}$ ($m = 4$). They are widely separated and one of them can be masked so as to transmit only one wavelength. In this way, it is possible to filter a particular wavelength out of a white light beam. Such a structure is known as an

interference filter.* Interference filters using this principle can be obtained by modern vacuum deposition techniques. A thin metallic film (usually, of aluminum or silver) is deposited on a substrate (generally, a glass plate) by vacuum deposition techniques. Then a thin layer of a dielectric material such as cryolite ($3\text{NaF}\cdot\text{AlF}_3$) is deposited over this. This structure is again covered by another metallic film (see Fig. 14.14). To protect this film structure from any damage, another glass plate is placed over it. Thus a Fabry–Perot structure is formed between the two glass plates. By varying the thickness of the dielectric film, one can filter out any particular wavelength. However, the filtered light will have a finite width, that is, it will have a narrow spectrum sharply peaked about one wavelength. The sharpness of the transmitted spectrum is determined by the resolving power of the formed Fabry–Perot structure, and hence by the reflectivity of the surfaces. The larger the reflectivity, the narrower is the transmitted spectrum. But it is not possible to increase the thickness of the metallic films indefinitely as absorption will reduce the intensity of the transmitted light. To overcome this difficulty, metallic films are replaced by all dielectric structures.



Fig. 14.14 The interference filter.

In an all-dielectric structure, layers of dielectric materials of appropriate refractive indices are deposited. It was shown in Chapter 13 how dielectric films can be used to enhance the reflectivity of a surface. If, on a glass-plate, a $\lambda/4$ thick film of a dielectric material whose refractive index is more than that of glass, is deposited, the reflectivity of the glass plate increases. Larger the difference between the refractive indices, greater will be the reflectivity. The materials generally used in interference filters are titanium oxide ($n = 2.8$) or zinc sulphide ($n = 2.3$). To obtain interference filters, a $\lambda/4$ thick film of titanium oxide is deposited on a glass substrate. Then a thin layer of dielectric material with lower refractive index (such as cryolite or magnesium fluoride) is deposited. On this is again deposited a $\lambda/4$ thick layer of a material of higher refractive index. To increase the reflectivity, multi-layer structures of alternate higher and lower refractive index materials are used. In this way, it is possible to achieve a reflectivity of more than 90% for any particular wavelength (see Sec. 13.6 for a more detailed account). Thus if the incident wave is polychromatic (like white light), the reflected light may have a high degree of monochromaticity.

SUMMARY

- If a plane wave falls on a plane parallel film, then the beam would undergo multiple reflections at the two surfaces and a large number of beams of successively diminishing amplitude will emerge on both sides of the plate. These beams (on either side) interfere to produce an interference pattern at infinity. If the reflectivity R at each surface is close to unity, then the fringes so formed are much sharper than those by two beam interference and, therefore, the interferometers involving multiple beam interference have a high resolving power and hence find applications in high resolution spectroscopy. The transmittivity of such a film is given by

$$T = \frac{1}{1 + F \sin^2 \frac{\delta}{2}}$$

where $F = \frac{4R}{(1-R)^2}$ is known as coefficient of Finesse and

$$\delta = \frac{4\pi n_2 h \cos \theta_2}{\lambda_0}$$

represents the phase difference (between two consecutive waves emanating from the film) due to the additional path traversed by the beam in the film; θ_2 is the angle of refraction inside the film (of refractive index n_2), h the film thickness and λ_0 is the free space wavelength. The transmittivity $T = 1$ when $\delta = 2m\pi$; $m = 1, 2, 3, \dots$ For $R = 1$, the value of F is very large and the transmission resonances become very sharp. This is the principle used in the Fabry–Perot interferometer which is characterized by a high resolving power.

PROBLEMS

- Calculate the resolving power of a Fabry–Perot interferometer made of reflecting surfaces of reflectivity 0.85 and separated by a distance 1 mm at $\lambda = 4880 \text{ \AA}$.
- Calculate the minimum spacing between the plates of a Fabry–Perot interferometer which would resolve two lines with $\Delta\lambda = 0.1 \text{ \AA}$ at $\lambda = 6000 \text{ \AA}$. Assume the reflectivity to be 0.8.
- Consider a monochromatic beam of wavelength $6,000 \text{ \AA}$ incident (from an extended source) on a Fabry–Perot etalon with $n_2 = 1$, $h = 1 \text{ cm}$ and $F = 200$. Concentric rings are observed on the focal plane of a lens of focal length 20 cm

*The Fabry–Perot structure also behaves as a resonator and supports the oscillation of what are known as modes.

- (a) Calculate the reflectivity of each mirror.
 (b) Calculate the radii of the first four bright rings. What will be the corresponding values of m ?
 (c) Calculate the angular width of each ring where the intensity falls by half and the corresponding FWHM (in mm) of each ring.
- 14.4** Consider now two wavelengths 6000 \AA and 5999.9 \AA incident on a Fabry–Perot etalon with the same parameters as given in the previous problem. Calculate the radii of the first three bright rings corresponding to each wavelength. What will be the corresponding values of m ? Will the lines be resolved?
- 14.5** Consider a monochromatic beam of wavelength 6000 \AA incident normally on a scanning Fabry–Perot interferometer with $n_2 = 1$ and $F = 400$. The distance between the two mirrors is written as $h = h_0 + x$. With $h_0 = 10 \text{ cm}$, calculate
- (a) The first three values of x for which we will have unit transmittivity and the corresponding values of m .
 (b) Also calculate the FWHM Δh for which the transmittivity will be half.
 (c) What would be the value of Δh if F was 200?
- [Ans. (a) $x \approx 200 \text{ nm}$ ($m = 333334$), 500 nm ($m = 333335$); (b) $\Delta h \approx 9.5 \text{ nm}$].
- 14.6** In continuation of Problem 14.5, consider now two wavelengths $\lambda_0 (= 6,000 \text{ \AA})$ and $\lambda_0 + \Delta\lambda$ incident normally on the Fabry–Perot interferometer with $n_2 = 1$, $F = 400$ and $h_0 = 10 \text{ cm}$. What will be the value of $\Delta\lambda$ so that $T = \frac{1}{2}$ occurs at the same value of h for both the wavelengths?
- 14.7** Consider a laser beam incident normally on the Fabry–Perot interferometer as shown in Fig. 14.15.
- (a) Assume $h_0 = 0.1 \text{ m}$, $c = 3 \times 10^8 \text{ m/s}$, $\nu = \nu_0 = 5 \times 10^{14} \text{ s}^{-1}$. Plot T as a function of x ($-100 \text{ nm} < x < 400 \text{ nm}$) for $F = 200$ and $F = 1000$.
 (b) Show that if $\nu = (\nu_0 \pm p \text{ 1500 MHz})$; $p = 1, 2, \dots$ we will have the same T vs. x curve; 1500 MHz is known as the free spectral range (FSR). What will be the corresponding values of δ ?

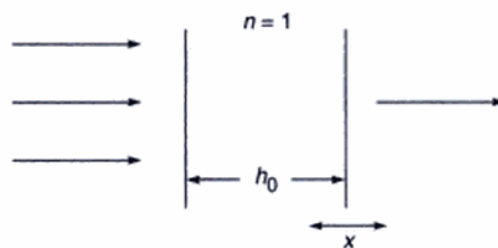


Fig. 14.15

REFERENCES AND SUGGESTED READINGS

1. R. Baierlein, *Newton to Einstein: the Trail of Light*, Cambridge University Press, 1992.
2. P. Baumeister and G. Pincus, 'Optical interference coatings', *Scientific American*, Vol. 223, 59, December, 1970.
3. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, 1975.
4. M. Cagnet, M. Francon and S. Mallick, *Atlas of Optical Phenomena*, Springer-Verlag, Berlin, 1971.
5. R. W. Ditchburn, *Light*, Academic Press, London, 1976.
6. M. Francon, *Modern Applications of Physical Optics*, Interscience, New York, 1963.
7. M. Francon, *Optical Interferometry*, Academic Press, New York, 1966.
8. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill Book Co., New York, 1976.
9. C. Lin, 'Optical communications: Single mode optical fiber transmission systems', *Optoelectronic Technology and Lightwave Communications Systems*, Ed. C. Lin, Van Nostrand Reinhold, New York, 1989.
10. W. H. Steel, *Interferometry*, Cambridge University Press, Cambridge, London, 1967.
11. S. Tolansky, *Multiple Beam Interferometry of Surfaces and Films*, Oxford University Press, London, 1948.
12. S. Tolansky, *An Introduction to Interferometry*, Longmans Green and Co., London, 1955.

Chapter 15

Coherence

Light which is capable of interference is called 'coherent', and it is evident that in order to yield many interference fringes, it must be very monochromatic. Coherence is conveniently measured by the path difference between two rays of the same source, by which they can differ while still giving observable interference contrast. This is called the coherence length... Lord Rayleigh and Albert Michelson were the first to understand that it is a reciprocal measure of the spectroscopic line width. Michelson used it for ingenious methods of spectral analysis and for the measurement of the diameter of stars.

— Dennis Gabor in his Nobel Lecture on Holography, December 11, 1971

15.1 INTRODUCTION

In earlier chapters on interference we had assumed that the displacement associated with a wave remained sinusoidal for all values of time. Thus the displacement (which we denote by E) was assumed to be given by

$$E = A \cos(kx - \omega t + \phi)$$

The above equation predicts that at any value of x , the displacement is sinusoidal for $-\infty < t < \infty$. For example, at $x = 0$ we have [see Fig. 15.1(a)].

$$E = A \cos(\omega t - \phi), \quad -\infty < t < \infty \quad (1)$$

Obviously this corresponds to an idealised situation because the radiation from an ordinary light source consists of finite size wavetrains, a typical variation of which is shown in Fig. 15.1(b). Since we will be considering only light waves, the quantity E represents the electric field associated with the light wave. Now, in Fig. 15.1(b), τ_c represents the average duration of the wavetrains, i.e., the electric field remains sinusoidal for times of the order of τ_c . Thus, at a given point, the electric fields at times t and $t + \Delta t$ will, in general, have a definite phase relationship if $\Delta t \ll \tau_c$ and will (almost) never have any phase relationship if $\Delta t \gg \tau_c$. The time duration τ_c is known as the coherence time of the source and the field is said to remain coherent for times $\sim \tau_c$. The length of the wavetrain, given by

$$L = c \tau_c \quad (2)$$

(where c is the speed of light in free space) is referred to as coherence length. For example, for the neon line ($\lambda = 6328 \text{ \AA}$), $\tau_c \sim 10^{-10}$ sec and for the red cadmium line

($\lambda = 6438 \text{ \AA}$), $\tau_c \sim 10^{-9}$ sec; the corresponding coherence lengths are 3 cm and 30 cm respectively. The finite value of the coherence time τ_c could be due to many factors; for example, if a radiating atom undergoes collision with another atom, then the wavetrain undergoes an abrupt phase shift of the type shown in Fig. 15.1(b). The finite coherence time could also be on account of the random motion of atoms or due to the fact that an atom has a finite lifetime in the energy level from which it drops to the lower energy level while radiating.*

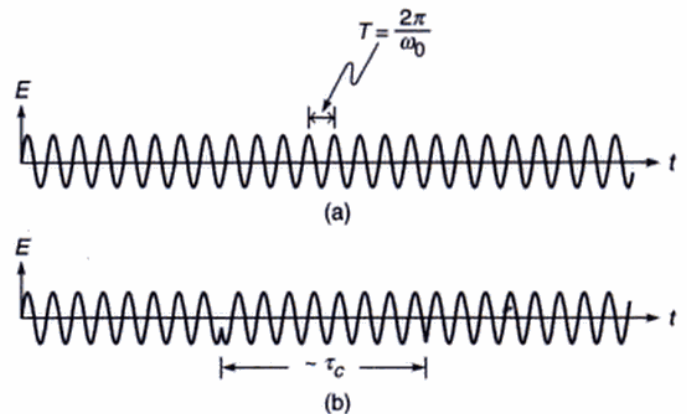


Fig. 15.1 (a) For a perfectly monochromatic beam, the displacement remains sinusoidal for $-\infty < t < +\infty$. (b) For an actual source, a definite phase relationship exists for times of the order of τ_c , which is known as the temporal coherence of the beam. For $\nu \sim 5 \times 10^{14}$ Hz and $\tau_c \sim 10^{-10}$ sec, one has about 50,000 oscillations in the time τ_c .

*For more details, See Ref. 15.

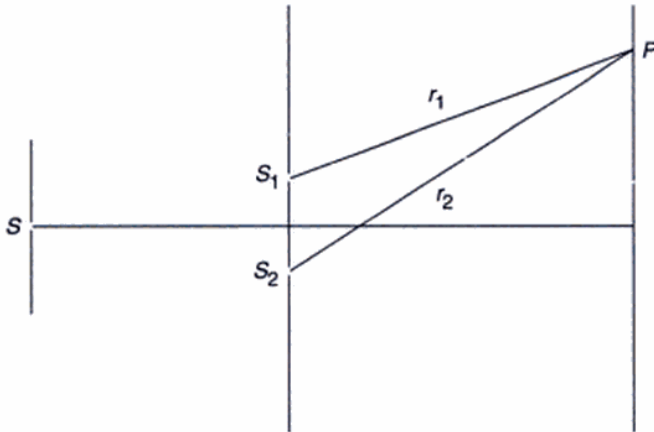


Fig. 15.2 Young's double-hole experiment. The interference pattern observed around the point P at time t is due to the superposition of waves emanating from S_1 and S_2 at times $t - \frac{r_1}{c}$ and $t - \frac{r_2}{c}$ respectively; thus interference fringes of good contrast will be observed at P if $(r_2 - r_1)/c \ll \tau_c$.

In order to understand the concept of coherence time (or of coherence length) we consider Young's double hole experiment as shown in Fig. 15.2; the interference pattern produced by this experimental arrangement was discussed in considerable detail in Sec. 12.4. Now, the interference pattern observed around the point P at time t is due to the superposition of waves emanating from S_1 and S_2 at times $t - r_1/c$ and $t - r_2/c$ respectively, where r_1 and r_2 are the distances S_1P and S_2P respectively. Obviously, if

$$\frac{r_2 - r_1}{c} \ll \tau_c$$

then the waves arriving at P from S_1 and S_2 will have a definite phase relationship and an interference pattern of good contrast will be obtained. On the other hand, if the path difference $(r_2 - r_1)$ is large enough such that

$$\frac{r_2 - r_1}{c} \gg \tau_c$$

then the waves arriving at P from S_1 and S_2 will have no fixed phase relationship and no interference pattern will be observed. Thus the central fringe (for which $r_1 = r_2$) will, in general, have a good contrast and as we move towards higher order fringes the contrast of the fringes will gradually become poorer. This point is discussed in greater detail in Sec. 15.7.

We next consider the Michelson interferometer experiment (see Sec. 13.10). A light beam falls on a beam splitter

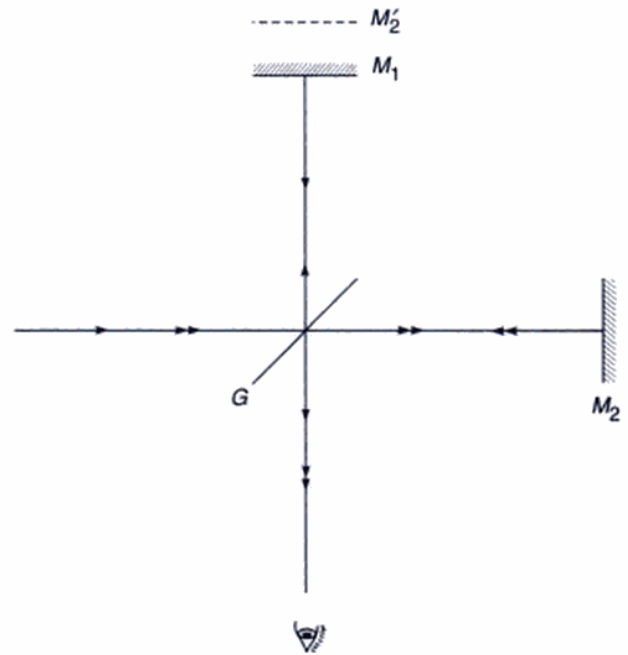


Fig. 15.3 The Michelson interferometer arrangement. G represents the beam splitter. M_2' represents the image of M_2 as formed by G .

G (which is usually a partially silvered plate) and the waves reflected from the mirrors M_1 and M_2 interfere (see Fig. 15.3). Let M_2' represent the image of the mirror M_2 (formed by the plate G) as seen by the eye. If the distance M_1M_2' is denoted by d , then the beam which gets reflected by mirror M_2 travels an additional path equal to $2d$. Thus, the beam reflected from M_1 interferes with the beam reflected by M_2 which had originated $2d/c$ seconds earlier.

If the distance d is such that

$$\frac{2d}{c} \ll \tau_c$$

then a definite phase relationship exists between the two beams and well-defined interference fringes are observed. On the other hand, if

$$\frac{2d}{c} \gg \tau_c$$

then, in general, there is no definite phase relationship between the two beams and no interference pattern is observed. It may be mentioned that there is no definite distance at which the interference pattern disappears; as the distance increases, the contrast of the fringes becomes gradually poorer and eventually the fringe system disappears. For the neon line ($\lambda = 6328 \text{ \AA}$), the disappearance occurs when the path difference is about a few centimetres giving $\tau_c \sim 10^{-10}$ sec. On the other hand for the red cadmium line ($\lambda = 6438 \text{ \AA}$), the coherence length is of the order of 30 cm giving $\tau_c \sim 10^{-9}$ sec.

The coherence time for a laser beam is usually much large in comparison to ordinary light sources. Indeed, for helium–neon laser, coherence times as large as 50 milliseconds have been obtained⁹; this would imply a coherence length of 15,000 km! Commercially available helium–neon lasers have $\tau_c \sim 50$ nsec implying coherence lengths of about 15 m. Thus using such a laser beam, high contrast interference fringes can be obtained even for a path difference of a few metres.

In order to demonstrate the large coherence length of the laser beam we consider an experimental arrangement shown in Fig. 15.4. A parallel beam of light is incident normally on a pair of circular holes. The Fraunhofer diffraction pattern is observed on the focal plane of a convex lens. We first use a helium neon laser beam, the resulting interference pattern is shown in Fig. 15.5(a) which is simply the product of the Airy pattern and the interference pattern produced by two point sources.* We next introduce a $\frac{1}{2}$ mm thick glass plate in front of one of the circular holes; there is almost no change in the interference pattern as can be seen from Fig. 15.5(b). Clearly, the extra path introduced by the plate $[(\mu - 1)t]$, see Sec. 12.10] is very small in comparison to the coherence length associated with the laser beam. If we repeat the experiment with a collimated mercury arc beam (Fig. 15.6), we would find that with the introduction of the glass plate the interference pattern disappears. This implies that the extra path length introduced by the glass plate is so large that there is no definite phase relationship between the waves arriving on the screen from the two circular apertures.

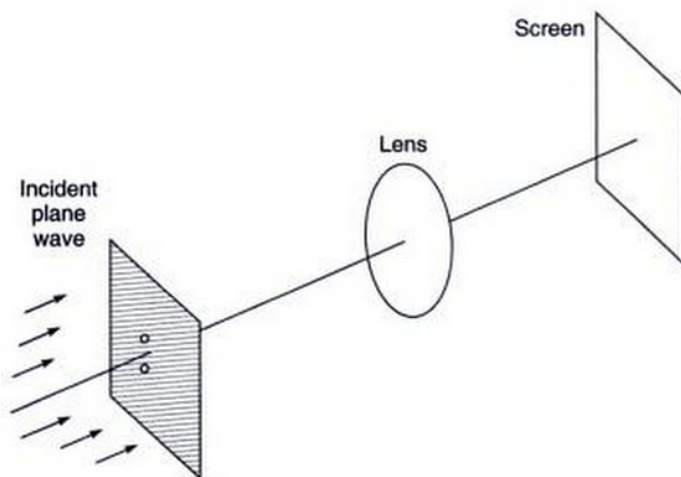


Fig. 15.4 A parallel beam of light is incident normally on a pair of circular holes and the Fraunhofer diffraction pattern is observed on the focal plane of a convex lens.

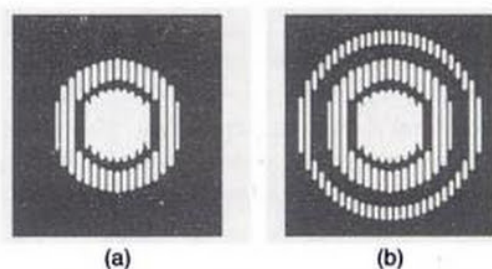


Fig. 15.5 (a) The interference pattern produced for the arrangement shown in Fig. 15.4 using a Helium–Neon laser beam; (b) The interference pattern produced by the same arrangement with a $\frac{1}{2}$ mm thick glass plate in front of one of the holes. [This and the following figure have been adapted from Ref. 16; the author came across the photographs in Ref. 8].

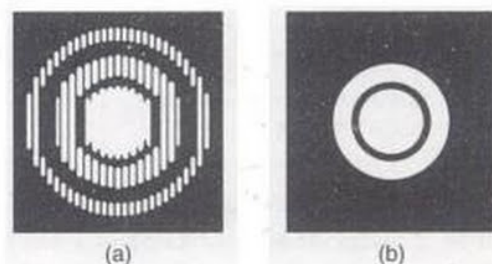


Fig. 15.6 (a) The interference pattern produced for the arrangement shown in Fig. 15.4 using a collimated mercury arc. (b) The interference pattern is washed out when $\frac{1}{2}$ mm thick glass plate is introduced in front of one of the holes.

15.2 THE LINEWIDTH

In the Michelson interferometer experiment discussed in the previous section, the decrease in the contrast of the fringes can also be interpreted as being due to the fact that the source is not emitting at a single frequency but over a narrow band of frequencies. When the path difference between the two interfering beams is zero or very small, the different wavelength components produce fringes superimposed on one another and the fringe contrast is good. On the other hand, when the path difference is increased, different wavelength components produce fringe patterns which are slightly displaced with respect to one another, and the fringe contrast becomes poorer. One can equally well say that the

*We may compare this with the double-slit Fraunhofer diffraction pattern which is simply the product of the single-slit diffraction pattern and the interference pattern produced by two point sources (see Sec. 16.5).

poor fringe visibility for a large optical path difference is due to the non-monochromaticity of the light source.

The equivalence of the above two approaches can be easily understood if we consider the Michelson interferometer experiment using two closely spaced wavelengths λ_1 and λ_2 . Indeed in Sec. 13.10 we had shown that for two closely spaced wavelengths λ_1 and λ_2 (like the D_1 and D_2 lines of sodium), the interference pattern will disappear if

$$\frac{2d}{\lambda_2} - \frac{2d}{\lambda_1} = \frac{1}{2} \quad (3)$$

where $2d$ represents the path difference between the two beams. Thus

$$2d = \frac{\lambda_1 \lambda_2}{2(\lambda_1 - \lambda_2)} \approx \frac{\lambda^2}{2(\lambda_1 - \lambda_2)} \quad (4)$$

Instead of two discrete wavelengths, if we assume that the beam consists of all wavelengths lying between λ and $\lambda + \Delta\lambda$, then the interference pattern produced by the wavelengths λ and $\lambda + \frac{1}{2}\Delta\lambda$ will disappear if

$$2d = \frac{\lambda^2}{2\left(\frac{1}{2}\Delta\lambda\right)} = \frac{\lambda^2}{\Delta\lambda} \quad (5)$$

Further, for each wavelength lying between λ and $\lambda + \frac{1}{2}\Delta\lambda$, there will be a corresponding wavelength (lying between $\lambda + \frac{1}{2}\Delta\lambda$ and $\lambda + \Delta\lambda$) such that the minima of one falls on the maxima of the other, making the fringes disappear. Thus, for

$$2d \geq \frac{\lambda^2}{\Delta\lambda} \quad (6)$$

the contrast of the interference fringes will be extremely poor. We may rewrite the above equation in the form

$$\Delta\lambda \geq \frac{\lambda^2}{2d} \quad (7)$$

implying that if the contrast of the interference fringes becomes very poor when the path difference is $\sim d$, then the spectral width of the source would be $\sim \lambda^2/2d$.

Now, in Sec. 15.1 we had observed that if the path difference exceeds the coherence length L , the fringes are not observed. From the above discussion it therefore follows that the spectral width of the source $\Delta\lambda$, will be given by

$$\Delta\lambda \sim \frac{\lambda^2}{L} = \frac{\lambda^2}{c\tau_c} \quad (8)$$

Thus the temporal coherence τ_c of the beam is directly related to the spectral width $\Delta\lambda$. For example, for the red cadmium line, $\lambda = 6438 \text{ \AA}$, $L \approx 30 \text{ cm}$ ($\tau_c \approx 10^{-9} \text{ sec}$) giving

$$\begin{aligned} \Delta\lambda &\sim \frac{\lambda^2}{c\tau_c} = \frac{(6438 \times 10^{-5})^2}{3 \times 10^{10} \times 10^{-9}} \\ &\sim 0.01 \text{ \AA} \end{aligned}$$

For the sodium line, $\lambda \approx 5890 \text{ \AA}$, $L \approx 3 \text{ cm}$ ($\tau_c \approx 10^{-10} \text{ sec}$) and $\Delta\lambda \sim 0.1 \text{ \AA}$. Further, since $\nu = c/\lambda$, the frequency spread $\Delta\nu$ of a line would be

$$\Delta\nu \sim \frac{c}{\lambda^2} \Delta\lambda \sim \frac{c}{L} \quad (9)$$

where we have disregarded the sign. Since $\tau_c = L/c$, we obtain

$$\Delta\nu \sim \frac{1}{\tau_c} \quad (10)$$

Thus the frequency spread of a spectral line is of the order of the inverse of the coherence time. For example, for the yellow line of sodium ($\lambda = 5890 \text{ \AA}$),

$$\begin{aligned} \tau_c \sim 10^{-10} \text{ s} &\Rightarrow \Delta\nu \sim 10^{10} \text{ Hz} \\ \nu = \frac{c}{\lambda} &= \frac{3 \times 10^{10}}{5890 \times 10^{-5}} \approx 5 \times 10^{14} \text{ Hz} \end{aligned}$$

we get

$$\frac{\Delta\nu}{\nu} \sim \frac{10^{10}}{5 \times 10^{14}} = 2 \times 10^{-5}$$

The quantity $\Delta\nu/\nu$ represents the monochromaticity (or the spectral purity) of the source and one can see that even for an ordinary light source it is very small. For a commercially available laser beam, $\tau_c \approx 50 \text{ nsec}$ implying $\Delta\nu/\nu \sim 4 \times 10^{-8}$. The fact that the finite coherence time is directly related to the spectral width of the source can also be seen using Fourier analysis; this is discussed in Sec. 15.6.

15.3 THE SPATIAL COHERENCE

Till now we have considered the coherence of two fields arriving at a particular point in space from a point source through two different optical paths. In this section we will discuss the coherence properties of the field associated with the finite dimension of the source.

We consider the Young's double-hole experiment with the point source S being equidistant from S_1 and S_2 [see Fig. 15.7(a)]. We assume S to be nearly monochromatic so that it produces interference fringes of good contrast on the screen PP' . The point O on the screen is such that $S_1O = S_2O$. Clearly, the point source S will produce an intensity maximum around the point O . We next consider another similar source S' at a distance l from S . We assume that the

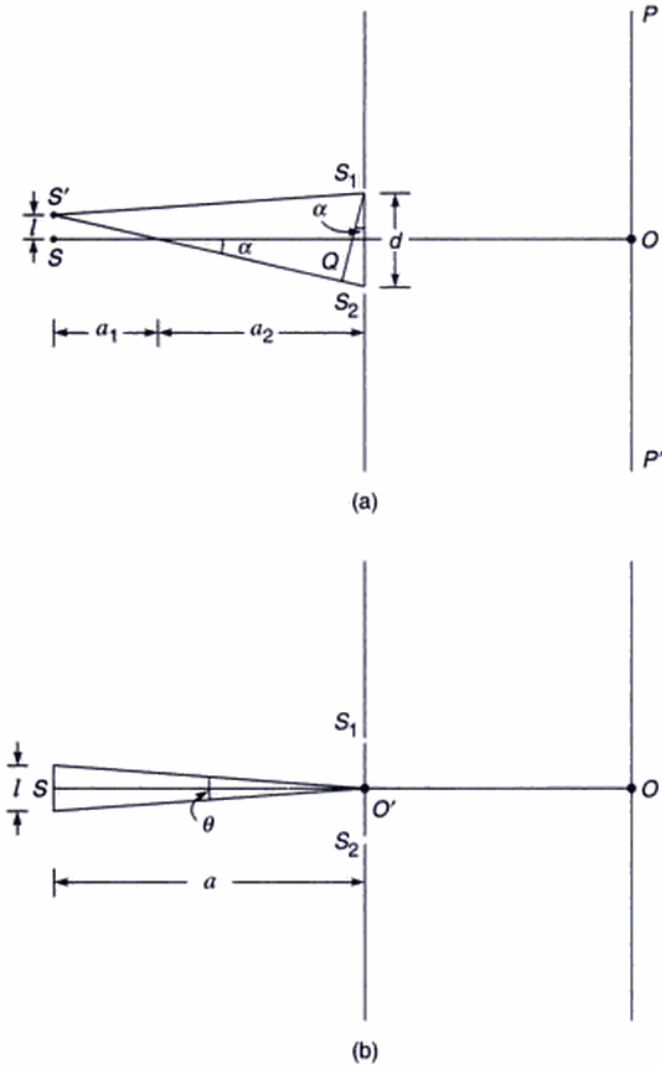


Fig. 15.7 (a) Young's double-hole interference experiment with two independent point sources S and S' . (b) The same experiment with an extended source.

waves from S and S' have no definite phase relationship. Thus the interference pattern observed on the screen PP' will be a superposition of the intensity distributions of the interference patterns formed due to S and S' (see Sec. 15.5). If the separation l is slowly increased from zero, the contrast of the fringes on the screen PP' becomes poorer because of the fact that the interference pattern produced by S' is slightly shifted from that produced by S . Clearly, if

$$S'S_2 - S'S_1 = \frac{\lambda}{2} \quad (11)$$

the minima of the interference pattern produced by S will fall on the maxima of the interference pattern produced by

S' and no fringe pattern will be observed. It can be easily seen that

$$S'S_2 = \left[a^2 + \left(\frac{d}{2} + l \right)^2 \right]^{1/2} \approx a + \frac{1}{2a} \left(\frac{d}{2} + l \right)^2$$

and

$$S'S_1 = \left[a^2 + \left(\frac{d}{2} - l \right)^2 \right]^{1/2} \approx a + \frac{1}{2a} \left(\frac{d}{2} - l \right)^2$$

where

$$a = a_1 + a_2$$

and we have assumed $a \gg d, l$. Thus

$$S'S_2 - S'S_1 \approx \frac{ld}{a}$$

Thus for the fringes to disappear, we must have

$$\frac{\lambda}{2} = S'S_2 - S'S_1 \approx ld/a$$

or

$$l \approx \frac{\lambda a}{2d}$$

Now, if we have an extended incoherent source whose linear dimension is $\sim \lambda a/d$ then for every point on the source, there is a point at a distance of $\lambda a/2d$ which produces fringes which are shifted by half a fringe width. Therefore the interference pattern will not be observed. Thus for an extended incoherent source, interference fringes of good contrast will be observed only when

$$l \ll \frac{\lambda a}{d} \quad (12)$$

Now, if θ is the angle subtended by the source at the slits [see Fig. 15.7(b)] then $\theta \approx l/a$ and the above condition for obtaining fringes of good contrast takes the form

$$d \ll \frac{\lambda}{\theta} \quad (13)$$

On the other hand, if

$$d \sim \frac{\lambda}{\theta} \quad (14)$$

the fringes will be of very poor contrast. Indeed, a more rigorous diffraction theory tells us that the interference pattern disappears when*

$$d = 1.22 \frac{\lambda}{\theta}, 2.25 \frac{\lambda}{\theta}, 3.24 \frac{\lambda}{\theta}, \dots \quad (15)$$

Thus as the separation of the pinholes is increased from zero, the interference fringes disappear when $d = 1.22\lambda/\theta$;

*See, for example, Ref. 7, Sec. 5.5

if d is further increased the fringes reappear with relatively poor contrast and they are washed out again when $d = 2.25\lambda/\theta$, and so on. The distance

$$l_w = \lambda/\theta \quad (\text{Lateral coherence width}) \quad (16)$$

gives the distance over which the beam may be assumed to be spatially coherent and is referred to as the *lateral coherence width*.

Example 15.1 On the surface of the earth, the sun subtends an angle of about $32'$. Assume sunlight to be falling normally on a double-hole arrangement of the type shown in Fig. 15.7 and that there is a filter in front of S_1S_2 so that light corresponding to $\lambda \approx 5000 \text{ \AA}$ is incident on S_1S_2 . What should be the separation between S_1 and S_2 so that fringes of good contrast are observed on the screen?

Solution:

$$\theta \approx 32' = \frac{32\pi}{180 \times 60} \text{ rad} \approx 0.01 \text{ rad}$$

Thus the lateral coherence length

$$l_w \approx \frac{5 \times 10^{-5}}{10^{-2}} = 0.005 \text{ cm}$$

Therefore if the pin holes are separated by a distance which is small compared to 0.005 cm, interference fringes of good contrast should be observed.

15.4 MICHELSON STELLAR INTERFEROMETER

Using the concept of spatial coherence, Michelson developed an ingenious method for determining the angular diameter of stars. The method is based on the result that for a distant circular source, the interference fringes will disappear if the distance between the pinholes S_1 and S_2 (see Fig. 15.8) is given by [see Eq. (15)]:

$$d = 1.22 \frac{\lambda}{\theta} \quad (17)$$

where θ is the angle subtended by the circular source as shown in Fig. 15.8. For a star whose angular diameter is 10^{-7} radians, the distance d for which the fringes will disappear would be

$$d \sim \frac{1.22 \times 5 \times 10^{-5}}{10^{-7}} \approx 600 \text{ cm}$$

where we have assumed $\lambda \approx 5000 \text{ \AA}$. Obviously, for such a large value of d , the fringe width will become extremely small. Further, one has to use a big lens, which is not only difficult to make, but only a small portion of which will be used. In order to overcome this difficulty, Michelson used

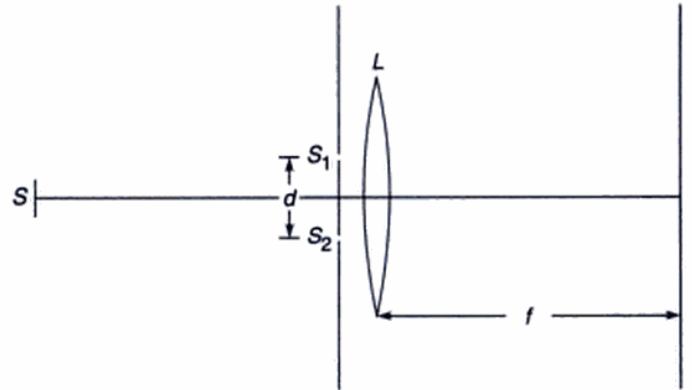


Fig. 15.8 S is a source of certain spatial extent; S_1 and S_2 are two slits separated by a distance d which can be varied. The fringes are observed on the focal plane of the lens L .

two movable mirrors M_1 and M_2 as shown in Fig. 15.9, and thus he effectively got a large value of d . The apparatus is known as Michelson's stellar interferometer. In a typical experiment the first disappearance occurred when the distance M_1M_2 was about 24 feet, which gave

$$\theta \approx \frac{1.22 \times 5 \times 10^{-5}}{24 \times 12 \times 2.54} \text{ radians} \approx 0.02''$$

for the angular diameter of the star. This star is known as Arcturus. From the known distance of the star, one can estimate that the diameter of the star is about 27 times that of the sun.

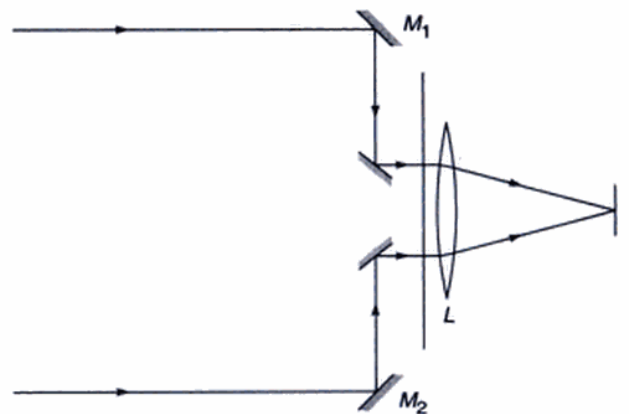


Fig. 15.9 Michelson's stellar interferometer.

We should point out that a laser beam is spatially coherent across the entire beam. Thus, if a laser beam is allowed to fall directly on a double-slit arrangement (see Fig. 15.10), then as long as the beam falls on both the slits, a clear interference pattern is observed on the screen. This shows that the laser beam is spatially coherent across the entire wavefront.

beats are so slow that one can see the wall get bright and dark".

15.6 COHERENCE TIME AND LINEWIDTH VIA FOURIER ANALYSIS

That the frequency spread of a line is of the order of the inverse of the coherence time [see Eq. (10)] can also be shown by Fourier analysis. As an example, we consider a sinusoidal displacement of duration τ_c . Thus we may write

$$\begin{aligned}\psi(x=0, t) &= ae^{i\omega_0 t} & |t| < \frac{1}{2}\tau_c \\ &= 0 & |t| > \frac{1}{2}\tau_c\end{aligned}\quad (28)$$

We will assume that τ_c is long enough so that the disturbance consists of many oscillations. For example, for a 2-nsec pulse corresponding to $\nu_0 \approx 5 \times 10^{14}$ Hz, the number of oscillations will be $5 \times 10^{14} \times 2 \times 10^{-9} = 10^6$, i.e. the pulse will consist of about a million oscillations!

Now, while discussing the Fourier transform theory (see Sec. 7.4), we had shown that for a time-dependent function $f(t)$, if we define

$$F(\omega) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(t) e^{-i\omega t} dt \quad (29)$$

then

$$f(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} F(\omega) e^{i\omega t} d\omega \quad (30)$$

Replacing $f(t)$ by $\psi(x=0, t)$, we may write

$$\psi(x=0, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} A(\omega) e^{i\omega t} d\omega \quad (31)$$

The RHS represents a superposition of plane waves with $A(\omega)$ representing the amplitude* of the plane wave corresponding to the frequency ω . Equation (31) tells us that $\psi(x=0, t)$ is the Fourier transform of $A(\omega)$ and therefore using the inverse Fourier transform [see Eq. (29)] we get

$$\begin{aligned}A(\omega) &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} \psi(x=0, t) e^{-i\omega t} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\frac{1}{2}\tau_c}^{+\frac{1}{2}\tau_c} ae^{i(\omega_0 - \omega)t} dt \\ &= \left(\frac{2}{\pi}\right)^{1/2} a \left[\frac{\sin\left\{\frac{1}{2}(\omega - \omega_0)\tau_c\right\}}{(\omega - \omega_0)} \right] \\ &= \left(\frac{2}{\pi}\right)^{1/2} \frac{a}{\omega_0} \left[\frac{\sin\{\alpha(\Omega - 1)\}}{(\Omega - 1)} \right]\end{aligned}\quad (32)$$

where $\Omega \equiv \frac{\omega}{\omega_0}$ and $\alpha = \frac{1}{2} \omega_0 \tau_c$. In Fig. 15.16 we have plotted the function

$$\frac{\sin[\alpha(\Omega - 1)]}{(\Omega - 1)} \quad (33)$$

as a function of Ω for $\alpha = 200$. One can see that the function is sharply peaked at $\Omega = 1$ (where it has a value equal to α) and that the first zero on either side occurs at $\Omega = 1 \pm (\pi/\alpha)$.

*Notice that the integral appearing on the RHS of Eq. (30) is over negative values of ω also. However, the displacement (or the electric field) is the real part of ψ which is given by (omitting the $\sqrt{2\pi}$ factor):

$$\begin{aligned}E &= \text{Re} [\psi(x=0, t)] = \text{Re} \left[\int_{-\infty}^{\infty} |A(\omega)| e^{i(\omega t + \phi)} d\omega \right] \\ &= \int_{-\infty}^{\infty} |A(\omega)| \cos(\omega t + \phi) d\omega = \int_0^{\infty} |A(\omega)| \cos(\omega t + \phi) d\omega + \int_0^{\infty} |A(-\omega)| \cos(\omega t - \phi(-\omega)) d\omega\end{aligned}$$

where we have used the relation $A(\omega) = |A(\omega)| e^{i\phi}$. The above equation can always be written in the form

$$\int_0^{\infty} C(\omega) \cos[\omega t + \theta(\omega)] d\omega$$

Thus the amplitudes associated with the negative frequencies contribute essentially to the corresponding positive frequencies.

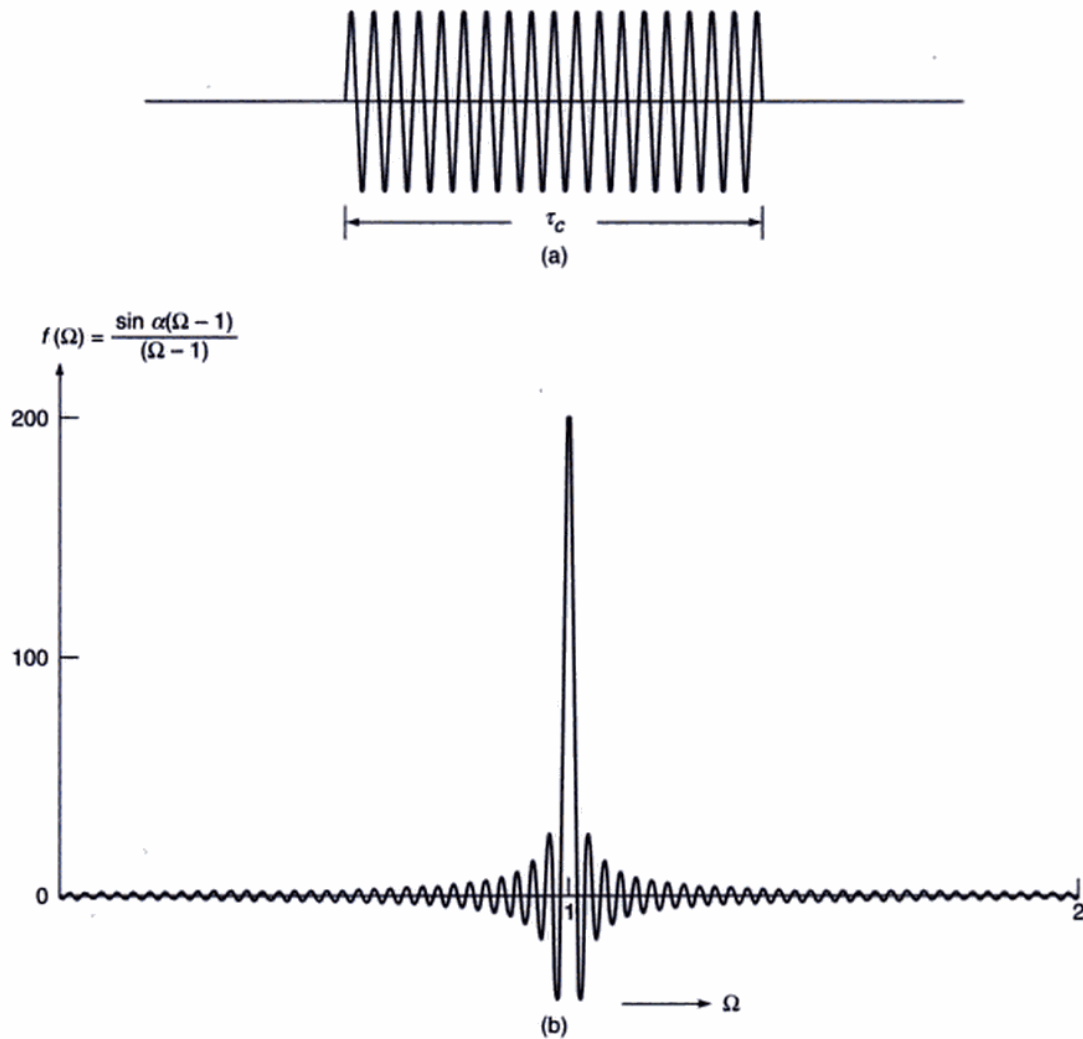


Fig. 15.16 (a) A sinusoidal displacement of duration τ_c ; (b) The variation of the function $[\sin (\Omega - 1)\alpha]/(\Omega - 1)$ as a function of Ω for $\alpha = 200$. Notice that the function is sharply peaked around $\Omega = 1$.

For larger values of α the function will become more sharply peaked; the width of the peak being given by

$$\Delta\Omega \left(= \frac{\Delta\omega}{\omega_0} \right) \sim \frac{\pi}{\alpha} \quad (34)$$

or

$$\Delta\omega \sim \frac{\pi\omega_0}{\alpha} \sim \frac{2\pi}{\tau_c}$$

Thus

$$\Delta\nu \sim \frac{1}{\tau_c} \quad (35)$$

consistent with Eq. (10). The above analysis shows that a wave having a coherence time $\sim \tau_c$ is essentially a superposition of harmonic waves having frequencies in the region

$$\nu_0 - \frac{1}{2} \Delta\nu \leq \nu \leq \nu_0 + \frac{1}{2} \Delta\nu \text{ where } \Delta\nu \sim 1/\tau_c.$$

We should mention that the condition expressed by Eq. (35) is quite general in the sense that it is valid for a pulse of arbitrary shape. For example, for a Gaussian pulse having a duration $\sim \tau_c$, the corresponding frequency spread will again be given by Eq. (35) [see Example 8.4].

15.7 COMPLEX DEGREE OF COHERENCE AND FRINGE VISIBILITY IN YOUNG'S DOUBLE-HOLE EXPERIMENT

In this section we will introduce the complex degree of coherence and will show how it can be related to the contrast of the fringes in the Young's double hole interference experiment. We refer back to Fig. 15.2. Let $\Psi_1(P, t)$ and $\Psi_2(P, t)$ represent the complex fields at the point P due to

Thus the visibility (or the contrast) of the fringes is a direct measure of $|\gamma_{12}|$. If $I_1 = I_2$ then $V = |\gamma_{12}|$. In the present case, since S , S_1 and S_2 have been assumed to be points $|\gamma_{12}|$ depends only on the temporal coherence of the beam. For $\tau \ll \tau_c$, $|\gamma_{12}|$ is very close to unity and the contrast of the fringes will be very good; for $\tau \gg \tau_c$, $|\gamma_{12}|$ will be close to zero and the contrast will be extremely poor.

It may be noted from Eq. (43) that for a perfectly monochromatic beam $|\gamma_{12}| = 1$ and $\alpha = \omega\tau = \omega(S_2P - S_1P)/c$. In general, it can be shown that $0 < |\gamma_{12}| < 1$; $|\gamma_{12}| = 0$ implies complete incoherence and $|\gamma_{12}| = 1$ implies complete coherence. In practice, if $|\gamma_{12}| > 0.88$, the light is said to be 'almost coherent'. Further, since

$$\langle \Psi^*(t + \tau)\Psi(t) \rangle = e^{i\omega\tau} \langle A(t + \tau)A(t)e^{i[\phi(t+\tau) - \phi(t)]} \rangle$$

and for a nearly monochromatic source $A(t)$ and $\phi(t)$ are already slowly varying functions of time, the quantity inside the angular brackets (on the RHS of the above equation) will not vary rapidly with τ . Thus, we may write

$$\gamma_{12} = |\gamma_{12}|e^{i\beta}e^{i\omega\tau} \quad (53)$$

where both $|\gamma_{12}|$ and β are slowly varying functions of

$$\tau = \frac{S_2P - S_1P}{c} \quad (54)$$

For a more detailed theory of spatial and temporal coherence, you may look in Refs. 2, 3, 7 and 20.

15.8 FOURIER TRANSFORM SPECTROSCOPY*

In the previous section we have shown that the contrast in an interference pattern depends on the relative magnitudes of the optical path difference Δ , vis a vis the coherence length of the source $L_c (= c\tau_c)$. For a given source, the contrast varies as the optical path difference Δ is varied, beginning from an extremely good contrast for $\Delta \ll L_c$ to a very poor contrast for $\Delta \gg L_c$. Indeed Fizeau in 1862 interpreted the periodic variation in contrast in Newton's rings under illumination with a sodium lamp as the lens is moved up, as being due to the presence of two lines separated by 6 \AA (see Example 13.4). Michelson in the years 1890–1900 performed various experiments with a number of spectral lines. Using the Michelson interferometer he measured visibility as a function of optical path difference and using a mechanical device he himself had built, he could obtain the spectra. It is the purpose of this section to show that from a knowledge of variation of intensity with optical path difference one can obtain the source spectral distribution by a Fourier transformation.

The use of the Michelson interferometer for spectroscopy was revived in the 1950s for application, specially for the relatively complex spectra in the infrared region.

We will derive expressions for the variation of visibility with optical path difference for a source having a certain spectral distribution and we will show that from the interference pattern one can obtain the spectral intensity distribution of the given source.

15.8.1 Principle of Fourier Transform Spectroscopy

Figure 15.17 shows the arrangement used in a Fourier transform spectrometer. Light from the given source is collimated and enters the Michelson interferometer and in the transmitted arm we measure the intensity at the focus of the lens as a function of the path difference Δ . Now, if a monochromatic beam of intensity I_0 is split into two beams (each of intensity $\frac{1}{2} I_0$) and are made to interfere, then the resultant intensity is given by

$$I = I_0(1 + \cos \delta) \quad (55)$$

where

$$\delta = \frac{2\pi}{\lambda} \Delta = \frac{2\pi\nu}{c} \Delta \quad (56)$$

represents the phase difference between the interfering beams, and in writing Eq. (55), we have used Eq. (30) of Chapter 12 with

$$I_1 = I_2 = \frac{1}{2} I_0$$

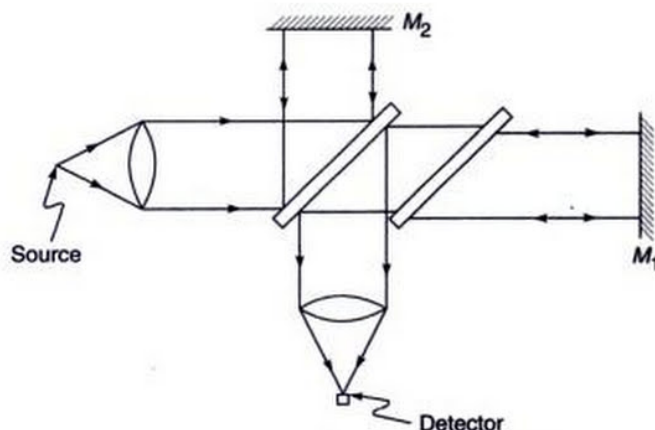


Fig. 15.17 The arrangement used in a Fourier transform spectrometer.

*Adapted from the unpublished lecture notes of Professor K. Thyagarajan.

Thus if $I(v) dv$ represents the intensity emitted by the source between v and $v + dv$ then the intensity at O lying between v and $v + dv$ is given by

$$I_t(v) dv = I(v) dv \left[1 + \cos \frac{2\pi v \Delta}{c} \right] \quad (57)$$

Hence, the total intensity at O corresponding to a path difference Δ is

$$\begin{aligned} I_t(\Delta) &= \int_0^{\infty} I_t(v) dv \\ &= \int_0^{\infty} I(v) dv + \int_0^{\infty} I(v) \cos \frac{2\pi v \Delta}{c} dv \end{aligned} \quad (58)$$

The quantity

$$I_T = \int_0^{\infty} I(v) dv = \frac{1}{2} I_t(0) \quad (59)$$

represents the total intensity of the source. We define normalized transmission as

$$\begin{aligned} \gamma(\Delta) &= \frac{I_t(\Delta) - I_T}{I_T} \\ &= \frac{1}{I_T} \int_0^{\infty} I(v) \cos \frac{2\pi v \Delta}{c} dv \end{aligned} \quad (60)$$

It is the quantity $I_t(\Delta)$ which is measured as a function of Δ from which $\gamma(\Delta)$ is evaluated. We first consider some examples giving explicit expressions for $I_t(\Delta)$ and $\gamma(\Delta)$ for some specific cases.

(i) Monochromatic Source

For a monochromatic source of intensity I_0 emitting at a frequency ν_0 , we have

$$I(v) dv = I_0 \delta(v - \nu_0) dv \quad (61)$$

where $\delta(v - \nu_0)$ represents the Dirac delta function. Hence

$$\begin{aligned} \gamma(\Delta) &= I_0 \frac{\int_0^{\infty} \delta(v - \nu_0) \cos \frac{2\pi v \Delta}{c} dv}{\int_0^{\infty} \delta(v - \nu_0) dv} \\ &= \cos \left(\frac{2\pi \nu_0 \Delta}{c} \right) \end{aligned} \quad (62)$$

and

$$I_t(\Delta) = I_0 \left(1 + \cos \frac{2\pi \nu_0 \Delta}{c} \right) \quad (63)$$

Hence $I_t(\Delta)$ and γ vary sinusoidally for all values of path difference Δ [see Figs 15.18(a) and (b)] implying that the coherence length of the source is infinite.

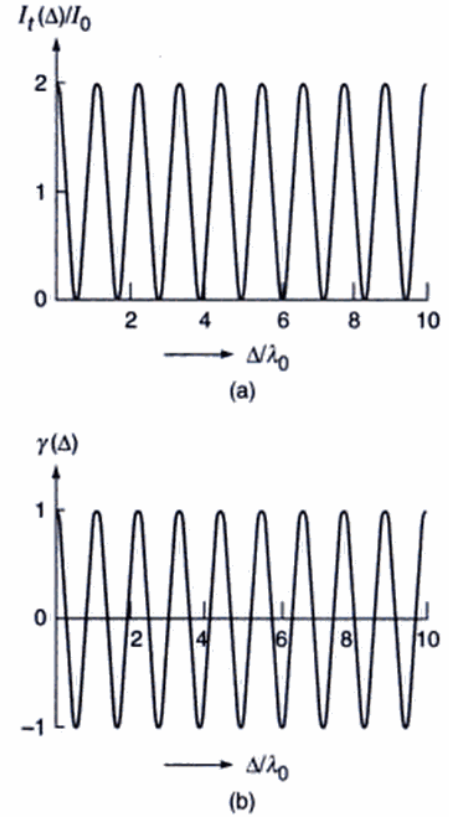


Fig. 15.18 (a) The variation of the total intensity at O as a function of the path difference Δ for a monochromatic source. (b) The corresponding cosinusoidal variation of $\gamma(\Delta)$ with Δ .

(ii) Source Emitting Two Monochromatic Lines

We now consider a source emitting two monochromatic lines at frequencies ν_1 and ν_2 , each characterized by an intensity $\frac{1}{2} I_0$. Thus

$$I(v) dv = \frac{1}{2} I_0 [\delta(v - \nu_1) + \delta(v - \nu_2)] \quad (64)$$

and

$$\begin{aligned} \gamma(\Delta) &= \frac{1}{2} \left[\int_0^{\infty} \delta(v - \nu_1) \cos \frac{2\pi v \Delta}{c} dv \right. \\ &\quad \left. + \int_0^{\infty} \delta(v - \nu_2) \cos \frac{2\pi v \Delta}{c} dv \right] \end{aligned}$$

$$\begin{aligned}
 &= \frac{1}{2} \left[\cos \frac{2\pi v_1 \Delta}{c} + \cos \frac{2\pi v_2 \Delta}{c} \right] \\
 &= \cos \left[2\pi \frac{(v_1 + v_2)}{2c} \Delta \right] \\
 &\times \cos \left[2\pi \frac{(v_1 - v_2)}{2c} \Delta \right] \quad (65)
 \end{aligned}$$

and

$$\begin{aligned}
 I_t(\Delta) &= I_0 \left\{ 1 + \cos \left[2\pi \frac{(v_1 - v_2)}{2c} \Delta \right] \right. \\
 &\quad \left. \times \cos \left[2\pi \frac{(v_1 + v_2)}{2c} \Delta \right] \right\} \quad (66)
 \end{aligned}$$

Such a variation of $I_t(\Delta)$ and $\gamma(\Delta)$ with Δ is shown in Fig. 15.19. From Eq. (65) we note that $\gamma(\Delta)$ corresponds to an amplitude modulated sinusoidal variation. The sinusoidal variation has a period

$$p = \frac{2c}{(v_1 + v_2)} = \frac{2\lambda_1 \lambda_2}{\lambda_1 + \lambda_2} \approx \lambda_0 \quad (67)$$

where $\lambda_0 (\approx \lambda_1 \approx \lambda_2)$ is the average wavelength. The modulation amplitude has zeroes at Δ values given by

$$2\pi \frac{(v_1 - v_2)}{2c} \Delta = \left(m + \frac{1}{2} \right) \pi$$

or

$$\Delta = \left(m + \frac{1}{2} \right) \frac{c}{v_1 - v_2} \quad (68)$$

Hence the minimum path difference at which the visibility vanishes is given by

$$\Delta_m = \frac{c}{2(v_1 - v_2)} = \frac{c}{2\delta\nu} \quad (69)$$

which corresponds to the coherence length of the source. Expressing $\delta\nu$ in terms of $\delta\lambda$, we have

$$L_c = \Delta_m = \frac{\lambda^2}{2\delta\lambda} \quad (70)$$

consistent with Eq. (2).

The difference in path difference between two consecutive positions of the disappearance of the fringes is $c/\delta\nu = \lambda^2/\delta\lambda$. As a simple consequence of this, we may consider the Newton's rings experiment with a sodium lamp. If we assume that the sodium lamp emits two discrete wavelengths λ_1 and λ_2 , then as we raise the convex lens above the glass plate we should have a periodic appearance of fringes as we had discussed in Example 13.3

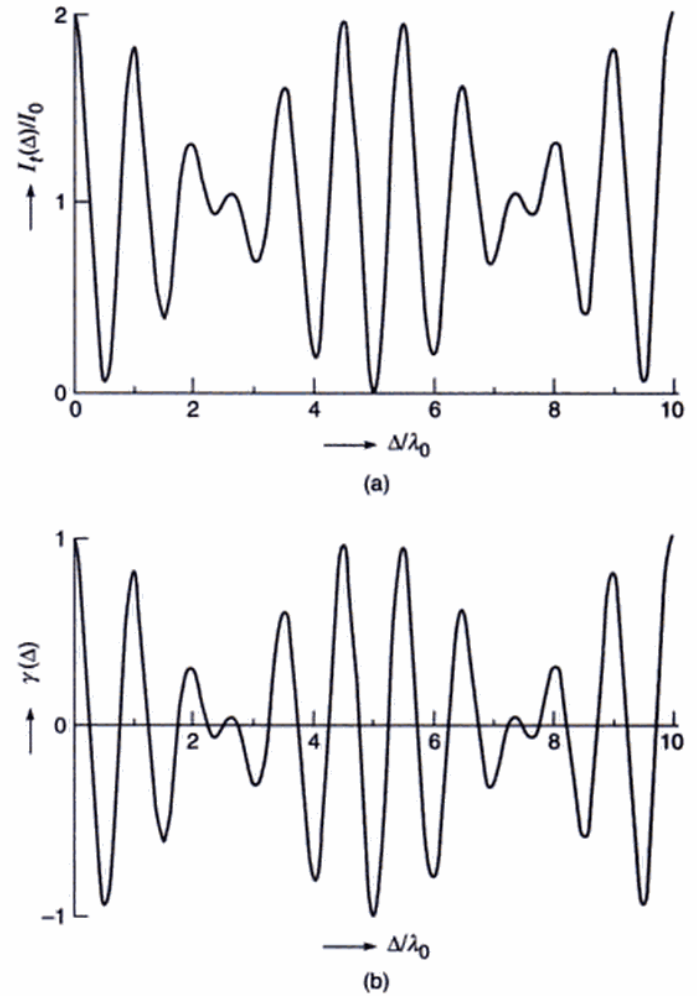


Fig. 15.19 (a) The variation of the total intensity at O as a function of the path difference Δ for a source emitting two monochromatic lines. (b) The corresponding variation of $\gamma(\Delta)$ with Δ .

15.8.2 Inversion to Recover $I(\nu)$ from $\gamma(\Delta)$

In an actual experiment, we measure $I_t(\Delta)$ and I_T . Thus Eq. (60) has to be inverted to obtain the source spectral distribution $I(\nu)$ from the measured $\gamma(\Delta)$. To do this we just multiply Eq. (60) by $\cos \frac{2\pi\nu'\Delta}{c}$ and integrate over Δ . Thus

$$\begin{aligned}
 &\int_0^\infty \gamma(\Delta) \cos \frac{2\pi\nu'\Delta}{c} d\Delta \\
 &= \frac{1}{I_T} \int_0^\infty d\Delta \int_0^\infty d\nu I(\nu) \cos \frac{2\pi\nu\Delta}{c} \cos \frac{2\pi\nu'\Delta}{c}
 \end{aligned}$$

$$= \frac{1}{I_T} \int_0^{\infty} d\nu I(\nu) \int_0^{\infty} \cos \frac{2\pi\nu\Delta}{c} \cos \frac{2\pi\nu'\Delta}{c} d\Delta$$

Now,

$$\begin{aligned} & \int_0^{\infty} \cos \frac{2\pi\nu\Delta}{c} \cos \frac{2\pi\nu'\Delta}{c} d\Delta \\ &= \frac{1}{2} \int_{-\infty}^{+\infty} \cos \frac{2\pi\nu\Delta}{c} \cos \frac{2\pi\nu'\Delta}{c} d\Delta \end{aligned}$$

since the integrand is an even function of Δ . Writing the two cosine terms in terms of exponentials and using

$$\int_{-\infty}^{+\infty} e^{\pm 2\pi i(\nu - \nu')\Delta/c} d\Delta = \delta\left(\frac{\nu - \nu'}{c}\right) = c\delta(\nu - \nu') \quad (71)$$

and

$$\int_{-\infty}^{+\infty} e^{\pm 2\pi i(\nu + \nu')\Delta/c} d\Delta = 0 \quad (72)$$

(since ν and ν' are positive), we obtain

$$\begin{aligned} \int_0^{\infty} \gamma(\Delta) \cos \frac{2\pi\nu'\Delta}{c} d\Delta &= \frac{c}{4I_T} \int_0^{\infty} \delta(\nu - \nu') I(\nu) d\nu \\ &= \frac{c}{4I_T} I(\nu') \end{aligned} \quad (73)$$

Hence

$$I(\nu) = \frac{4I_T}{c} \int_0^{\infty} \gamma(\Delta) \cos \frac{2\pi\nu\Delta}{c} d\Delta \quad (74)$$

Thus one can obtain the source spectral distribution $I(\nu)$ from the measured $\gamma(\Delta)$ just by a cosine transformation. Such an inversion from $\gamma(\Delta)$ to $I(\nu)$ is usually performed using a computer.

15.8.3 Resolution

From Eq. (74), it follows that to obtain $I(\nu)$ one must measure $\gamma(\Delta)$ for all values of path difference Δ lying between 0 and ∞ . Since in an actual experiment, there is a maximum limit to path differences that can be introduced, this maximum path difference determines the resolution obtainable in the estimated $I(\nu)$. To estimate the resolution, we consider a perfectly monochromatic beam of frequency ν_0 incident on the interferometer. We have seen that for such a case $\gamma(\Delta)$ varies with Δ as given by Eq. (62). Now in the experiment if Δ_m is the maximum path difference measured, then $\gamma(\Delta)$ would be

$$\begin{aligned} \gamma(\Delta) &= \cos \frac{2\pi\nu_0\Delta}{c} & 0 < \Delta < \Delta_m \\ &= 0 & \text{otherwise} \end{aligned} \quad (75)$$

Hence using Eq. (74), we have

$$\begin{aligned} I(\nu) &= \frac{4I_T}{c} \int_0^{\Delta_m} \cos\left(\frac{2\pi\nu_0\Delta}{c}\right) \cos\left(\frac{2\pi\nu\Delta}{c}\right) d\Delta \\ &= \frac{2I_T}{c} \int_0^{\Delta_m} \left[\cos \frac{2\pi(\nu + \nu_0)\Delta}{c} + \cos \frac{2\pi(\nu - \nu_0)\Delta}{c} \right] d\Delta \\ &= \frac{2I_T}{c} \left[\frac{\sin\left(\frac{2\pi(\nu + \nu_0)\Delta_m}{c}\right)}{\frac{2\pi}{c}(\nu + \nu_0)} + \frac{\sin\left(\frac{2\pi(\nu - \nu_0)\Delta_m}{c}\right)}{\frac{2\pi}{c}(\nu - \nu_0)} \right] \end{aligned}$$

Since ν and ν_0 are both positive and much much greater than c/Δ , the first term in the RHS within brackets is negligible and we obtain

$$I(\nu) \approx \frac{2I_T}{c} \left[\frac{\sin \frac{2\pi(\nu - \nu_0)\Delta_m}{c}}{\frac{2\pi}{c}(\nu - \nu_0)} \right] \quad (76)$$

The above estimated source spectrum is similar to that shown in Fig. 15.16. The spectrum is peaked at ν_0 and the first zero appears at

$$\frac{2\pi(\nu - \nu_0)}{c} \Delta_m = \pm\pi$$

or

$$\nu = \nu_0 \pm \frac{c}{2\Delta_m} \quad (77)$$

Thus although the incident beam is monochromatic, the inversion process gives us a finite spectral width due to a finite value of Δ_m .

If the incident source contains two frequencies, then we may use the Rayleigh criterion and define the minimum resolvable frequency separation to be the frequency width from the peak to the first zero in $I(\nu)$. Hence

$$\delta\nu = \frac{c}{2\Delta_m} \quad (78)$$

Hence, the larger the maximum path difference Δ_m over which γ is measured, the higher will be the resolution.

As an example if $\Delta_m = 5$ cm, then

$$\delta\nu = \frac{3 \times 10^{10}}{2 \times 5} = 3 \text{ GHz}$$

At $\lambda = 1 \mu\text{m}$, this corresponds to $\delta\lambda = 0.1 \text{ \AA}$.

PART 4

Diffraction

Chapters 16 and 17 cover the very important area of diffraction and discuss the principle behind topics like diffraction divergence of laser beams, resolving power of telescopes, laser focusing, spatial frequency filtering, X-ray diffraction etc. Chapter 18 is on holography giving the underlying principle and many applications. Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the principles of holography.

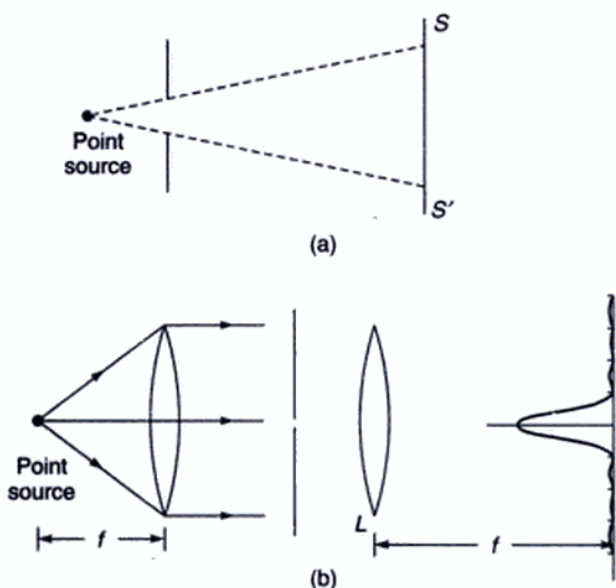


Fig. 16.2 (a) When either the source or the screen (or both) are at finite distances from the aperture, the diffraction pattern corresponds to the Fresnel class. (b) In the Fraunhofer class both the source and the screen are at infinity.

source and the screen to infinity because the first lens makes the light beam parallel and the second lens effectively makes the screen receive a parallel beam of light. It turns out that it is much easier to calculate the intensity distribution of a Fraunhofer diffraction pattern which we plan to do in this chapter. Further, the Fraunhofer diffraction pattern is not difficult to observe; all that one needs is an ordinary laboratory spectrometer; the collimator renders a parallel beam of light and the telescope receives parallel beams of light on its focal plane. The diffracting aperture is placed on the prism table. In the next chapter we will study the Fresnel class of diffraction and will discuss the transition from the Fresnel region to the Fraunhofer region.

16.2 SINGLE-SLIT DIFFRACTION PATTERN

We will first study the Fraunhofer diffraction pattern produced by an infinitely long slit of width b . A plane wave is assumed to fall normally on the slit and we wish to calculate the intensity distribution on the focal plane of the lens L [see Fig. 16.3(a)]. We assume that the slit consists of a large number of equally spaced point sources and that each point on the slit is a source of Huygens' secondary wavelets which interfere with the wavelets emanating from other points. Let the point sources be at A_1, A_2, A_3, \dots and let the distance

between two consecutive points be Δ [see Fig. 16.3(b)]. Thus, if the number of point sources be n , then

$$b = (n - 1)\Delta \quad (1)$$

We will now calculate the resultant field produced by these n sources at the point P , P being an arbitrary point (on the focal plane of the lens) receiving parallel rays making an angle θ with the normal to the slit [see Fig. 16.3(b)]. Since the slit actually consists of a continuous distribution of sources, we will, in the final expression, let n go to infinity and Δ go to zero such that $n\Delta$ tends to b .

Now, at the point P , the amplitudes of the disturbances reaching from A_1, A_2, \dots will be very nearly the same because the point P is at a distance which is very large in comparison to b . However, because of even slightly different path lengths to the point P , the field produced by A_1 will differ in phase from the field produced by A_2 .

For an incident plane wave, the points A_1, A_2, \dots are in phase and, therefore, the additional path traversed by the disturbance emanating from the point A_2 will be A_2A_2' where A_2' is the foot of the perpendicular drawn from A_1 on A_2B_2 . This follows from the fact that the optical paths A_1B_1P and $A_2'B_2P$ are the same. If the diffracted rays make an angle θ with the normal to the slit then the path difference would be

$$A_2A_2' = \Delta \sin \theta$$

The corresponding phase difference, ϕ , would be given by

$$\phi = \frac{2\pi}{\lambda} \Delta \sin \theta \quad (2)$$

Thus, if the field at the point P due to the disturbance emanating from the point A_1 is $a \cos \omega t$ then the field due to the disturbance emanating from A_2 would be $a \cos (\omega t - \phi)$. Now the difference in the phases of the disturbance reaching from the points A_2 and A_3 will also be ϕ and thus the resultant field at the point P would be given by

$$E = a[\cos \omega t + \cos (\omega t - \phi) + \dots + \cos (\omega t - (n - 1)\phi)] \quad (3)$$

where

$$\phi = \frac{2\pi}{\lambda} \Delta \sin \theta$$

Now, we had shown in Sec. 11.7 that

$$\begin{aligned} &\cos \omega t + \cos (\omega t - \phi) + \dots + \cos [\omega t - (n - 1)\phi] \\ &= \frac{\sin n\phi/2}{\sin \phi/2} \cos \left[\omega t - \frac{1}{2} (n - 1)\phi \right] \end{aligned} \quad (4)$$

Thus

$$E = E_0 \cos \left[\omega t - \frac{1}{2} (n - 1)\phi \right] \quad (5)$$

Thus

$$E = A \frac{\sin \beta}{\beta} \cos(\omega t - \beta) \quad (9)$$

The corresponding intensity distribution is given by

$$I = I_0 \frac{\sin^2 \beta}{\beta^2} \quad (10)$$

where I_0 represents the intensity at $\theta = 0$

16.2.1 Positions of Maxima and Minima

The variation of the intensity with β is shown in Fig. 16.4(a). It is obvious from Eq. (10) that the intensity is zero when

$$\beta = m\pi, m \neq 0 \quad (11)$$

[When $\beta = 0$, $\frac{\sin \beta}{\beta} = 1$ and $I = I_0$ which corresponds to the maximum of the intensity.] Substituting the value of β one obtains

$$b \sin \theta = m\lambda; m = \pm 1, \pm 2, \pm 3, \dots \text{ (minima)} \quad (12)$$

as the conditions for minima. The first minimum occurs at

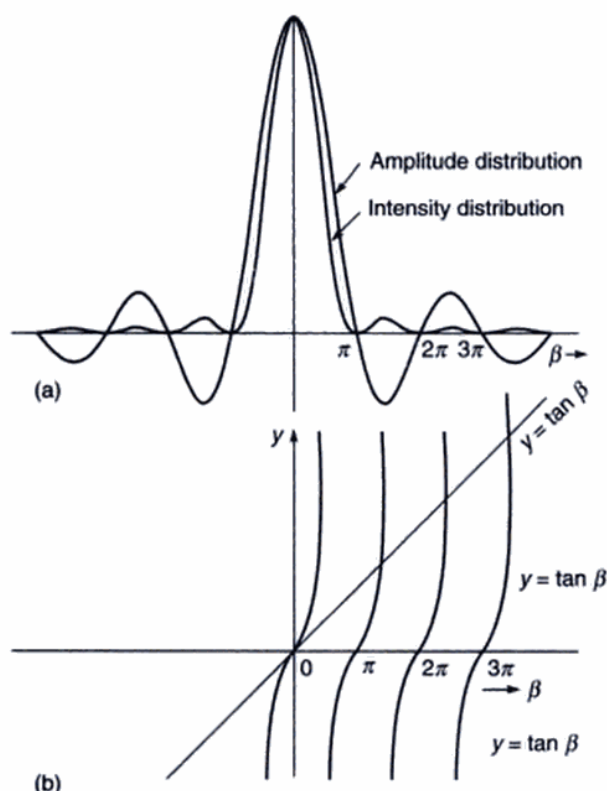


Fig. 16.4 (a) The intensity distribution corresponding to the single slit Fraunhofer diffraction pattern. (b) Graphical method for determining the roots of the equation $\tan \beta = \beta$.

$\theta = \pm \sin^{-1} \left(\frac{\lambda}{b} \right)$; the second minimum at $\theta = \pm \sin^{-1} \left(\frac{2\lambda}{b} \right)$, etc. Since $\sin \theta$ cannot exceed unity, the maximum value of m is the integer which is less than (and closest to) $\frac{b}{\lambda}$.

The positions of minima can directly be obtained by simple qualitative arguments. Let us consider the case $m = 1$. The angle θ satisfies the equation

$$b \sin \theta = \lambda \quad (13)$$

We divide the slit into two halves as shown in Fig. 16.5. Consider two points A and A' separated by a distance $b/2$. Clearly the path difference between the disturbances (reaching the point P) emanating from A and A' is $\frac{b}{2} \sin \theta$ which in this case is $\frac{\lambda}{2}$. The corresponding phase difference will be π and the resultant disturbance will be zero. Similarly, the disturbance from the point B will be cancelled by the disturbance reaching from the point B' . Thus the resultant disturbance due to the upper half of the slit will be cancelled by the disturbances reaching from the lower half and the resultant intensity will be zero. In a similar manner when

$$b \sin \theta = 2\lambda \quad (14)$$

we divide the slit into four parts; the first and second quarters cancelling each other and the third and fourth quarters cancelling each other. Similarly when $m = 3$, the slit is divided into six parts and so on.

In order to determine the positions of maxima, we differentiate Eq. (10) with respect to β and set it equal to zero. Thus

$$\frac{dI}{d\beta} = I_0 \left[\frac{2 \sin \beta \cos \beta}{\beta^2} - \frac{2 \sin^2 \beta}{\beta^3} \right] = 0$$

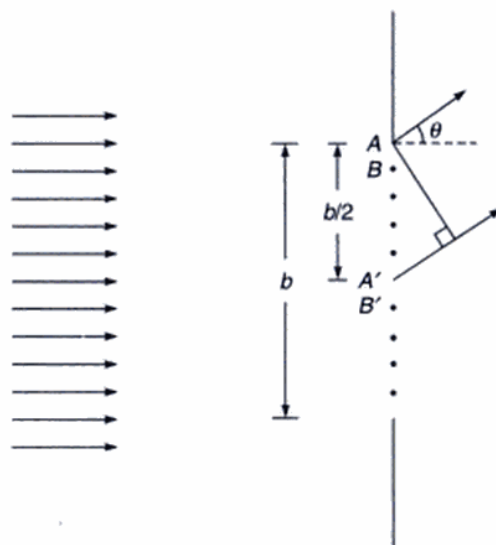


Fig. 16.5 The slit is divided into two halves for deriving the condition for the first minimum.

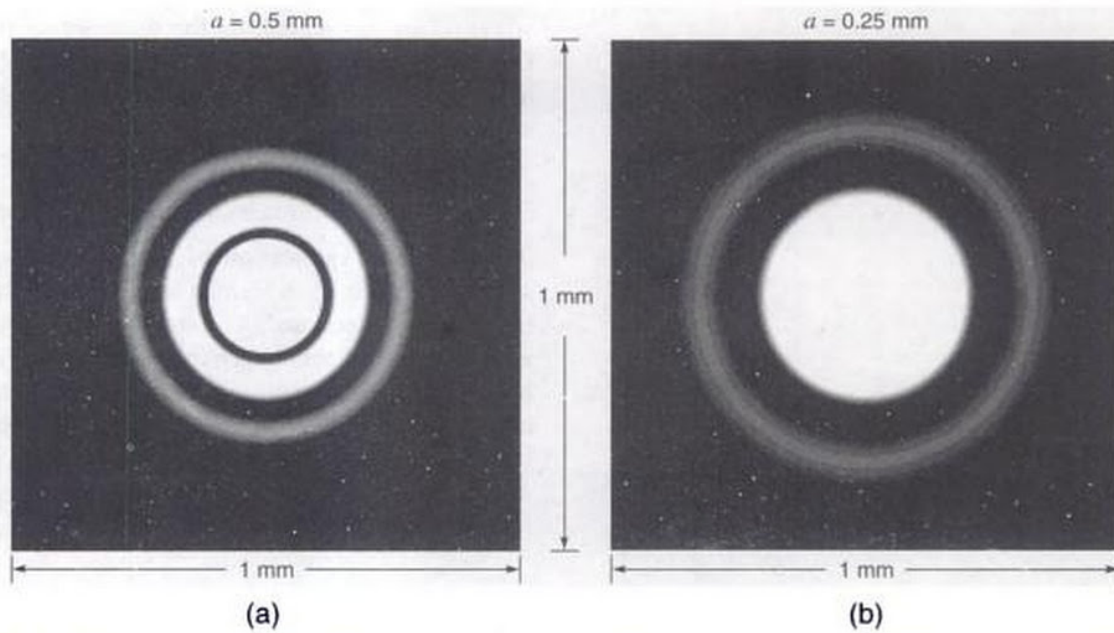


Fig. 16.9 Computer generated Airy patterns; (a) and (b) correspond to $a = 0.5$ mm and $a = 0.25$ mm respectively at the focal plane of a lens of focal length 20 cm ($\lambda = 0.5 \mu\text{m}$).

somewhat complicated (see Sec. 17.8.2); we give here the final result: the intensity distribution is given by

$$I = I_0 \left[\frac{2J_1(v)}{v} \right]^2 \quad (19)$$

where

$$v = \frac{2\pi}{\lambda} a \sin \theta \quad (20)$$

a being the radius of the circular aperture, λ the wavelength of light and θ the angle of diffraction; I_0 is the intensity at $\theta = 0$ (which represents the central maximum) and $J_1(v)$ is known as the Bessel function of the first order. On the focal plane of the convex lens

$$v = \frac{2\pi}{\lambda} a \frac{(x^2 + y^2)^{\frac{1}{2}}}{f} \quad (21)$$

where f is the focal length of the lens. For those not familiar with Bessel functions, we may mention that the variation of $J_1(v)$ is somewhat like a damped sine curve (see Fig. 16.10) and although $J_1(0) = 0$, we have

$$\lim_{v \rightarrow 0} \frac{2J_1(v)}{v} = 1$$

similar to the relation

$$\lim_{x \rightarrow 0} \frac{\sin x}{x} = 1$$

Other zeros of $J_1(v)$ occur at

$$v = 3.832, 7.016, 10.174, \dots$$

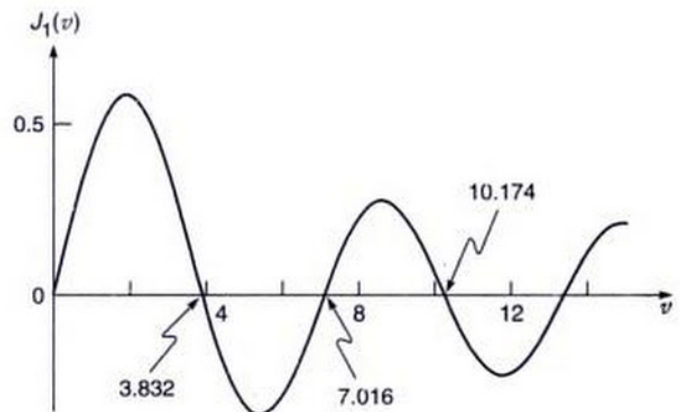


Fig. 16.10 The variation of $J_1(v)$ with v .

In Fig. 16.11 we have plotted the function

$$\left[\frac{2J_1(v)}{v} \right]^2$$

which represents the intensity distribution corresponding to the Airy pattern. Thus the successive dark rings in the Airy pattern (see Fig. 16.9) will correspond to

$$\begin{aligned} v &= \frac{2\pi}{\lambda} a \sin \theta \\ &= 3.832, 7.016, 10.174, \dots \end{aligned} \quad (22)$$

or

$$\sin \theta = \frac{3.832 \lambda}{2\pi a}, \frac{7.016 \lambda}{2\pi a}, \dots \quad (23)$$

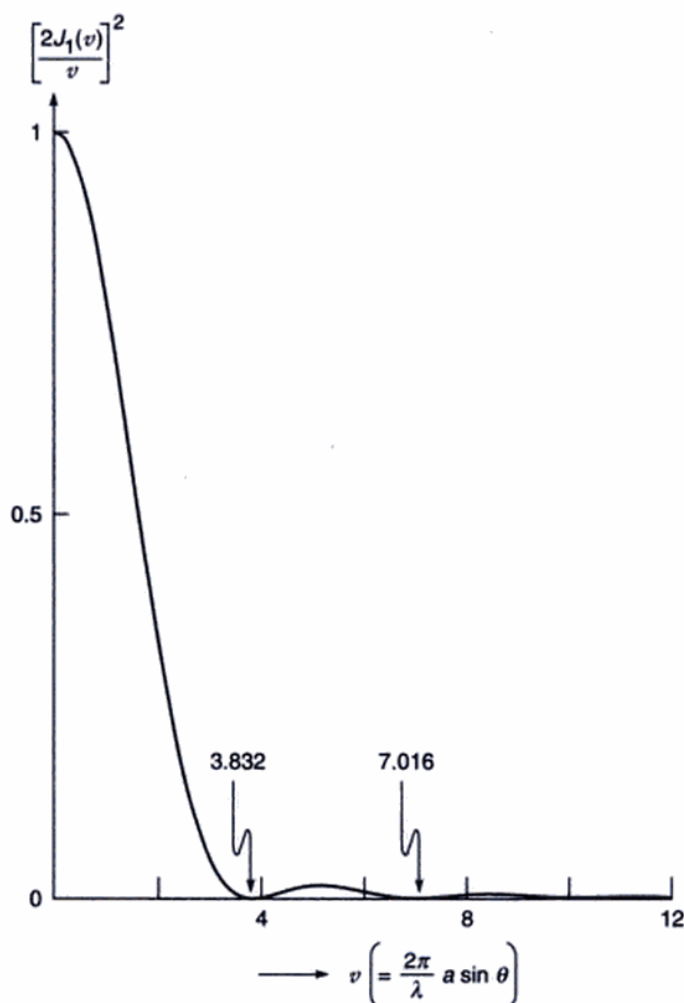


Fig. 16.11 The intensity variation associated with the Airy pattern.

If f represents the focal length of the convex lens, then the Radii of the dark rings

$$= f \tan \theta \approx \frac{3.832 \lambda f}{2\pi a}, \frac{7.016 \lambda f}{2\pi a}, \dots \quad (24)$$

where we have assumed θ to be small so that $\tan \theta \approx \sin \theta$. The Airy patterns shown in Figs 16.9(a) and (b) correspond to $a = 0.5$ mm and 0.25 mm respectively; both figures correspond to $\lambda = 5000\text{\AA}$ and $f = 20$ cm. Thus

Radius of the first dark ring ≈ 0.12 mm and 0.24 mm

corresponding to $a = 0.5$ mm and 0.25 mm respectively. Detailed mathematical analysis shows that about 84% of the energy is contained within the first dark ring (see Sec. 17.8.2); thus we may say that the angular spread of the beam is approximately given by

$$\Delta\theta \approx \frac{0.61 \lambda}{D} \approx \frac{\lambda}{D} \quad (25)$$

where $D (= 2a)$ represents the diameter of the aperture. Comparing Eqs. (18) and (25), we may say that the angular divergence associated with the diffraction pattern can be written in the following general form:

$$\Delta\theta \sim \frac{\lambda}{\text{Linear dimension of the aperture}} \quad (26)$$

An interesting application of the above phenomenon is shown in Fig. 16.12. A layman would expect that in order to obtain more directionality of sound waves, one should use a loudspeaker of small aperture as shown in Fig. 16.12(a); however, this will result in a greater diffraction divergence and only a small fraction of energy will reach the observer. On the other hand, if one uses a loudspeaker of larger diameter, greater directionality is achieved [see Fig. 16.12(b)].

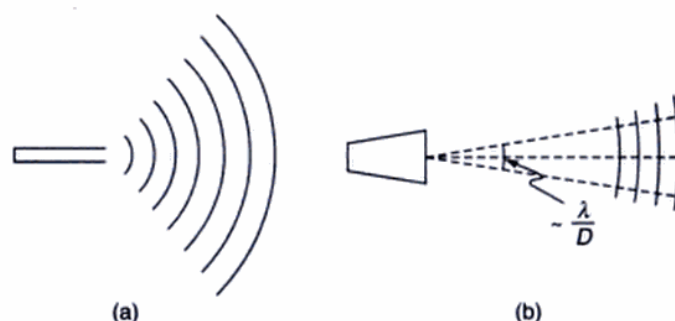


Fig. 16.12 The directionality of sound waves increases with increase in the diameter of the speaker.

Example 16.3 Calculate the radii of the first two dark rings of the Fraunhofer diffraction pattern produced by a circular aperture of radius 0.02 cm at the focal plane of a convex lens of focal length 20 cm. Assume $\lambda = 6 \times 10^{-5}$ cm.

Solution: The first dark ring occurs at

$$\theta \approx \sin \theta = \frac{1.22 \times 6 \times 10^{-5}}{2 \times 0.02} \approx 1.8 \times 10^{-3} \text{ radians}$$

Thus the radius of the first dark ring is

$$\approx 20 \times 1.8 \times 10^{-3} = 3.6 \times 10^{-2} \text{ cm}$$

Similarly, the radius of the second dark ring is

$$\approx 20 \times \frac{7.016 \times 6 \times 10^{-5}}{2\pi \times 0.02} \approx 6.7 \times 10^{-2} \text{ cm}$$

16.4 DIRECTIONALITY OF LASER BEAMS

An ordinary source of light (like a sodium lamp) radiates in all directions. On the other hand, the divergence of a laser beam is primarily due to diffraction effects. For most laser

by $f \approx 2.5$ cm. If the pupil diameter ($= 2a$) is taken to be 2 mm, then

$$\text{Area of the focused spot } A = \pi \left(\frac{\lambda_0 f}{a} \right)^2 \approx 7 \times 10^{-6} \text{ cm}^2$$

On the retina, the intensity will be approximately given by

$$I = \frac{P}{A} = \frac{2 \times 10^{-3} \text{ W}}{7 \times 10^{-10} \text{ m}^2} \approx 3 \times 10^6 \text{ W/m}^2$$

Such high intensities will damage the retina!!!! **So never look into a (seemingly innocent) low power laser beam.**

Example 16.6 We next consider a 3 MW laser beam ($\lambda_0 \approx 6 \times 10^{-5}$ cm and beam width $2a \approx 1$ cm) incident on a lens of focal length of 5 cm, then

Area of the focused spot

$$A = \pi \left(\frac{\lambda_0 f}{a} \right)^2 \approx 10^{-6} \text{ cm}^2 = 10^{-10} \text{ m}^2$$

On the focal plane of the lens, the intensity will be approximately given by

$$I = \frac{P}{A} = \frac{3 \times 10^6 \text{ W}}{10^{-10} \text{ m}^2} \approx 3 \times 10^{16} \text{ W/m}^2$$

Now, the intensity of the beam is related to the electric field amplitude E_0 through the following relation [see Eq. (78) of Chapter 20]

$$I = \frac{1}{2} \epsilon_0 c E_0^2 \quad (34)$$

where $\epsilon_0 \approx 8.854 \times 10^{-12}$ MKS units represents the dielectric permittivity of free space and $c \approx 3 \times 10^8$ m/s represents the speed of light in free space. Substituting $I \approx 3 \times 10^{16}$ W/m² in Eq. (34) we readily get

$$E_0 \approx 5 \times 10^9 \text{ V/m}$$

Such high electric fields results in the creation of spark in air (see Fig. 16.16). Thus laser beams (because of their high directionality) can be focused to extremely small regions producing very high intensities. Such high intensities lead to numerous industrial applications of the laser such as welding, hole drilling, cutting materials, etc.⁵

In the following two examples, we will calculate the intensities (at the retina of our eye) when we directly view a 500 W bulb or the sun (Caution: Never look into the sun; the retina will be damaged not only because of high intensities but also because of large ultraviolet content of the sunlight).

Example 16.7 We consider a 6 cm diameter incandescent source (like a 500 W bulb) at a distance of about 5 m from the eye (see Fig. 16.17). We assume the pupil diameter to be about 2 mm. Thus

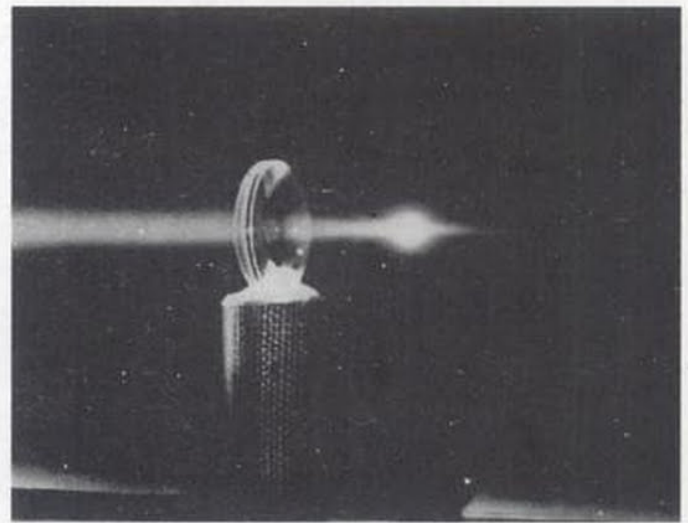


Fig. 16.16 Focusing of a 3 MW peak power pulsed ruby laser beam. At the focus, the electric field strengths are of the order of 10^9 V/m which results in the creation of a spark in the air. (Photograph courtesy Dr. R. W. Terhune).

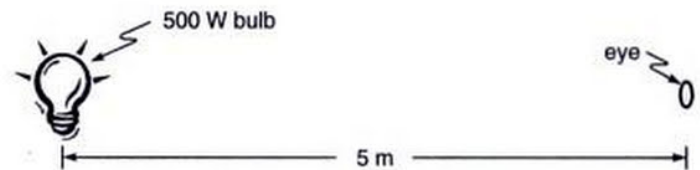


Fig. 16.17 A 500 W bulb at a distance of about 5 m from the eye.

$$\text{Area of the pupil of the eye} = \pi (1 \times 1) \text{ mm}^2 = 3 \times 10^{-6} \text{ m}^2$$

$$\text{Power entering eye} = (500 \text{ W}) \times \frac{\pi r^2}{4\pi R^2} \approx 5 \times 10^{-6} \text{ W}$$

$$\text{Radius of image} = \text{Radius of source} \times \text{demagnification}$$

$$\approx 3 \text{ cm} \times \frac{2.5}{500} \approx 1.5 \times 10^{-4} \text{ m}$$

where we have assumed the image to be formed at a distance of about 2.5 cm from the pupil of the eye. Thus,

The power density in image

$$= \frac{(5 \times 10^{-6} \text{ W})}{\pi (1.5 \times 10^{-4})^2 \text{ m}^2} \approx 70 \text{ W/m}^2$$

Example 16.8 We next calculate the intensity at the retina if we are directly looking at the sun (see Fig. 16.18). Now

$$\text{The intensity of solar energy on earth} \approx 1.35 \text{ kW/m}^2$$

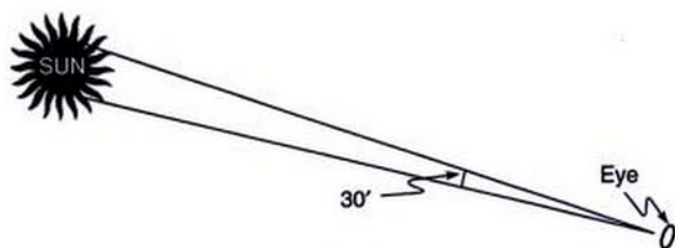


Fig. 16.18 If we look directly at the sun, intensities as high as 30 kW/m^2 are produced; this can damage the retina of the eye!

Thus the energy entering the eye

$$\approx 1.35 \times 10^3 \times \pi \times 10^{-6} \approx 4 \text{ mW}$$

The sun subtends about 0.5° on the earth. Thus

The radius of the image of the sun

$$\begin{aligned} &\approx 0.5 \times \frac{\pi}{180} \times 2.5 \approx 0.2 \text{ mm} \\ &= 2 \times 10^{-4} \text{ m} \end{aligned}$$

and

The power density in image

$$\begin{aligned} &\approx \frac{4 \times 10^{-3} \text{ W}}{\pi \times (2 \times 10^{-4})^2 \text{ m}^2} \\ &\approx 30 \text{ kW/m}^2 \end{aligned}$$

To summarize, a 2 mW diffraction limited laser beam incident on the eye can produce an intensity of about 10^6 W/m^2 at the retina—this would certainly damage the retina. Thus, whereas it is quite safe to look at a 500 W bulb, it is very dangerous to look directly into a 2 mW laser beam. Indeed, because a laser beam can be focused to very narrow areas, it has found important applications in areas like eye surgery, welding, etc.

From the above discussion it immediately follows that greater the radius of the beam, the smaller will be the size of the focused spot and hence greater will be the intensity at the focused spot. Indeed, one may use a beam expander (see Fig. 16.19) to produce a beam of greater size and hence a smaller focused spot size. However, after the focused spot, the beam would have a greater divergence and would therefore expand within a very short distance. One usually defines a *depth of focus* as the distance over which the intensity of the beam (on the axis) decreases by a certain factor of the value at the focal point. Thus a small focused spot would lead to a small depth of focus. We may mention here that the intensity distribution at the focal plane of the lens is given by Eq. (19) where the parameter v is given by

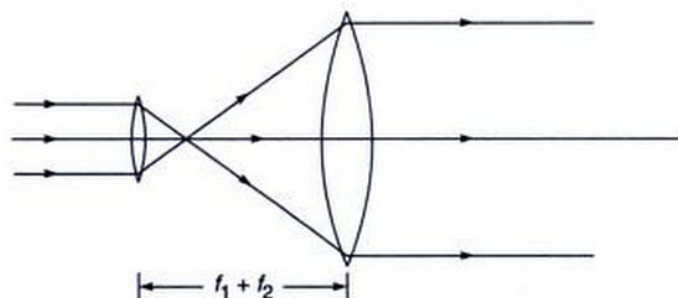


Fig. 16.19 Two convex lenses separated by a distance equal to the sum of their focal lengths act like a beam expander.

Eq. (21). On the other hand, the intensity along the axis is given by

$$I = I_0 \left[\frac{\sin(w/4)}{w/4} \right]^2 \quad (35)$$

where

$$w = \frac{2\pi}{\lambda} \left(\frac{a}{f} \right)^2 z \quad (36)$$

and $z = 0$ represents the focal plane.*

It can be readily seen that the intensity would drop by about 20% at

$$z \approx \pm 0.5 \lambda (f/a)^2 \quad (37)$$

which is usually referred to as the depth of the focus or focal tolerance. Notice that larger the value of a , smaller will be the focal tolerance. For $\lambda \approx 6 \times 10^{-5} \text{ cm}$, $f \approx 10 \text{ cm}$ and $a \approx 1 \text{ cm}$, the focal tolerance is about $3 \times 10^{-3} \text{ cm}$.

16.5 LIMIT OF RESOLUTION

Consider two point sources, such as stars (so that we can consider plane waves entering the aperture) being focused by a telescope objective of diameter D (see Fig. 16.20). As discussed in the previous section, the system can be thought of as being equivalent to a circular aperture of diameter D , followed by a converging lens of focal length f , as shown in Fig. 16.8. As such, each point source will produce its Airy pattern as schematically shown in Fig. 16.20. The diameters of the Airy rings will be determined by the diameter of the objective, its focal length and the wavelength of light (see Example 16.3).

In Fig. 16.20 the Airy patterns are shown to be quite far away from each other and, therefore, the two objects are said to be well resolved. Since the radius of the first ring is

*The derivation of the formulae has been given at many places—see, e.g., Section 5.5 of Ref. 6.

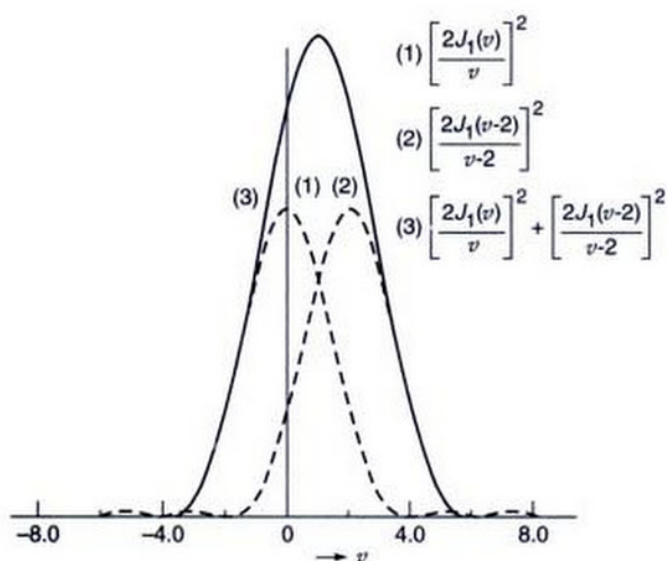


Fig. 16.22 The dashed curves correspond to the intensity distribution produced independently by two distant point objects having an angular separation of $2\lambda / \pi D$. The resultant intensity distribution (shown as a solid curve) has only one peak and hence the objects are unresolved.

$$\frac{1.22\lambda}{D} \times \text{focal length} = \frac{1.22 \times 6 \times 10^{-5}}{5} \times 30 \\ \approx 4.5 \times 10^{-4} \text{ cm}$$

It is immediately obvious that the larger the diameter of the objective, the better will be its resolving power. For example, the diameter of the largest telescope objective is about 80" and the corresponding angular separation of the objects that it can resolve is ≈ 0.07 sec. of arc. This very low limit of resolution is never achieved in ground based telescopes due to the turbulence of the atmosphere. However, a larger aperture still provides a larger light gathering power and hence the ability to see deeper in space.

It is of interest to note that if we assume that the angular resolution of the human eye is primarily due to diffraction effects then it will be given by

$$\Delta\theta \sim \frac{\lambda}{D} \approx \frac{6 \times 10^{-5}}{2 \times 10^{-1}} = 3 \times 10^{-4} \text{ rad.} \quad (40)$$

where we have assumed the pupil diameter to be 2 mm. Thus, at a distance of 20 m, the eye should be able to resolve two points which are separated by a distance

$$3 \times 10^{-4} \times 20 = 6 \times 10^{-3} \text{ m} = 6 \text{ mm}$$

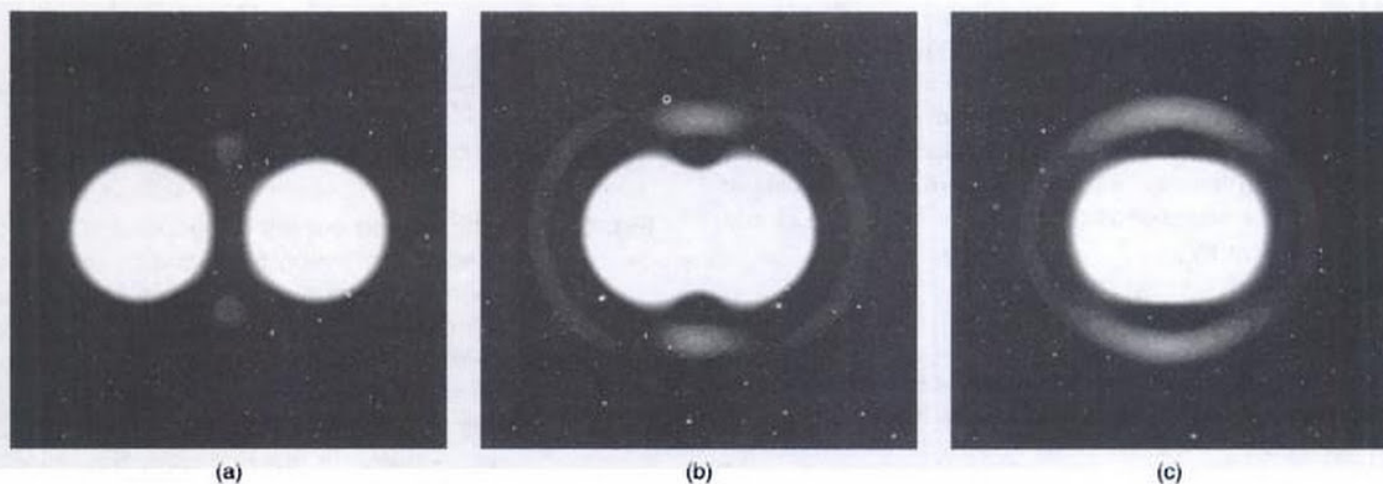


Fig. 16.23 Computer generated intensity distributions corresponding to two point sources when they are: (a) well resolved, (b) just resolved, and (c) unresolved.

minimum angular separation of two distant objects which can just be resolved will be

$$\frac{1.22\lambda}{D} = \frac{1.22 \times 6 \times 10^{-5}}{5} \approx 1.5 \times 10^{-5} \text{ radians}$$

Further, the radius of the first dark ring (of the Airy pattern) will be

One can indeed verify that this result is qualitatively valid by finding the distance at which the millimetre scale will become blurred.

In the above discussion we have assumed that the two object points produce identical (but displaced) Airy patterns. If that is not the case then the two central maxima will have different intensities; accordingly one has to set up

a modified criterion for the limit of resolution such that the two maxima stand out.

16.5.1 Resolving Power of a Microscope

We next consider the resolving power of a microscope objective of diameter D as shown in Fig. 16.24. Let P and Q represent two closely spaced self-luminous point objects which are to be viewed through the microscope. Assuming the absence of any geometrical aberrations, rays emanating from the points P and Q will produce spherical wavefronts (after refraction through the lens) which will form Airy patterns around their paraxial image points P' and Q' . For the points P and Q to be just resolved, the point Q' should lie on the first dark ring surrounding the point P' and therefore we must have

$$\sin \alpha' \approx \frac{1.22 \lambda}{D} = \frac{1.22 \lambda_0}{n'D} \quad (41)$$

where n and n' represent the refractive indices of the object and image spaces, λ_0 and $\lambda (= \lambda_0/n')$ represent the wavelength of light in free space and in the medium of refractive index n' respectively. The angle α' is defined in Fig. 16.24 and we have

$$\sin \alpha' \approx \frac{y'}{OP'} = \frac{y' \tan i'}{D/2} \approx \frac{y' \sin i'}{D/2} \quad (42)$$

where we have assumed $\sin i' \approx \tan i'$, this is justified since the image distance (OP') is large compared to D . Using Eqs (41) and (42), we get

$$y' \approx \frac{0.61 \lambda_0}{n' \sin i'}$$

If we now use the sine law $n'y' \sin i' = ny \sin i$ [see Eq. 39 of Chapter 3], we get

$$y \approx \frac{0.61 \lambda_0}{n \sin i} \quad (43)$$

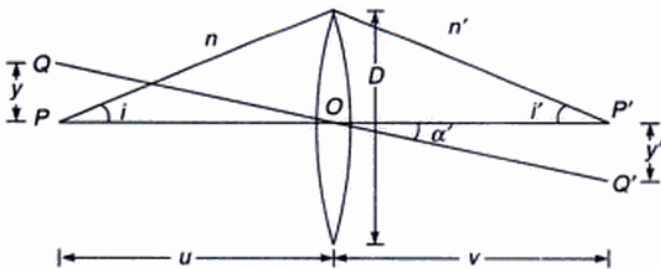


Fig. 16.24 The resolving power of a microscope objective.

which represents the smallest distance that the microscope can resolve. The quantity $n \sin i$ is the numerical aperture of the optical system and the resolving power increases with increase in the numerical aperture. It is for this reason that in some microscopes the space between the object and the objective is filled with an oil—and they are referred to as ‘oil immersion objectives’. Equation (43) also tells us that the resolving power increases with decrease in λ . As such, one often uses blue light (or even ultraviolet light) for the illumination of the object. For example, in an electron microscope the de Broglie wavelength of electrons accelerated to 100 keV is about 0.03×10^{-8} cm and therefore such a microscope has a very high resolving power.

In the above analysis, we have assumed that the two object points are self-luminous so that the intensities can be added up. However, in actual practice, the objects are illuminated by the same source and, therefore, in general, there is some phase relationship between the waves emanating from the two object points; for such a case the intensities will not be strictly additive (see Sec. 12.6), nevertheless Eq. (43) will give the correct order for the limit of resolution.

16.6 TWO-SLIT FRAUNHOFER DIFFRACTION PATTERN

In Section 16.3 we had studied the Fraunhofer diffraction pattern produced by a slit of width b and had found that the intensity distribution consisted of maxima and minima. In this section we will study the Fraunhofer diffraction pattern produced by two parallel slits (each of width b) separated by a distance d . We would find that the resultant intensity distribution is a product of the single-slit diffraction pattern and the interference pattern produced by two point sources separated by a distance d .

In order to calculate the diffraction pattern we use a method similar to that used for the case of a single slit and assume that the slits consist of a large number of equally spaced point sources and that each point on the slit is a source of Huygens’ secondary wavelets. Let the point sources be at A_1, A_2, A_3, \dots (in the first slit) and at B_1, B_2, B_3, \dots (in the second slit) [see Fig. 16.25]. As before, we assume that the distance between two consecutive points in either of the slits is Δ . If the diffracted rays make an angle θ with the normal to the plane of the slits, then the path difference between the disturbances reaching the point P from two consecutive points in a slit will be $\Delta \sin \theta$. The field produced by the first slit at the point P will, therefore, be given by [see Eq. (9)]

$$E_1 = A \frac{\sin \beta}{\beta} \cos (\omega t - \beta)$$

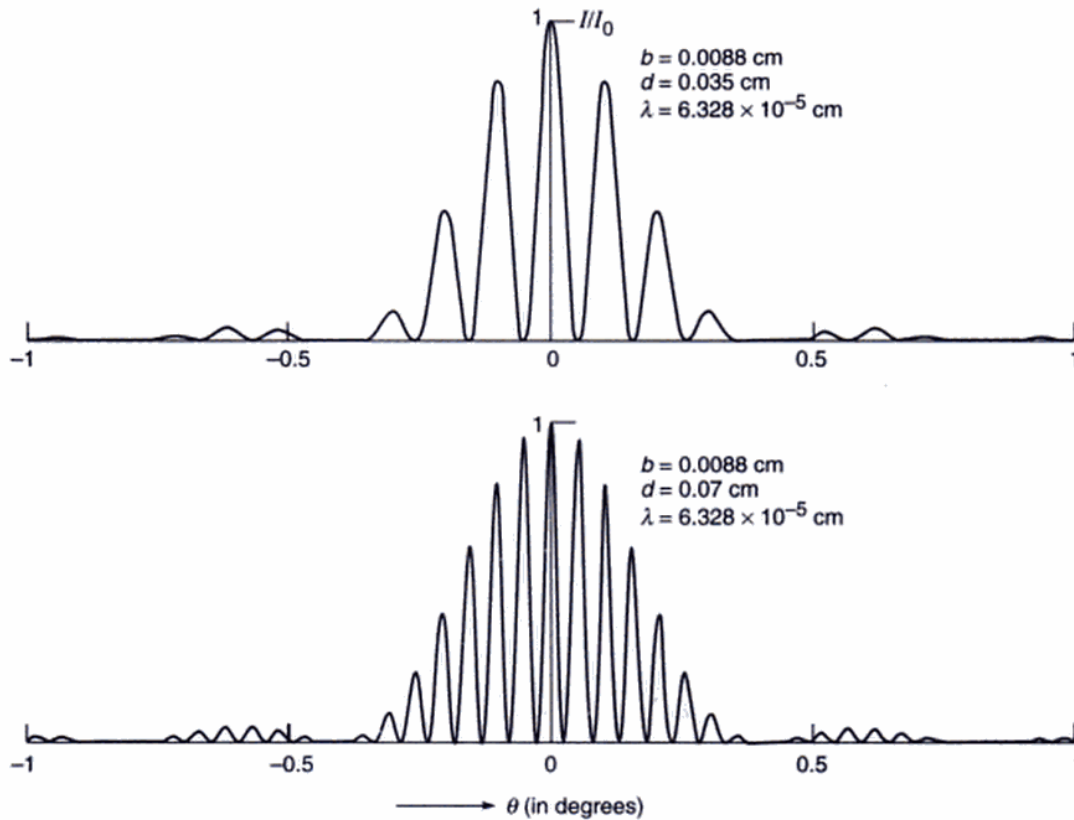


Fig. 16.27 The double-slit intensity distribution as predicted by Eq. (45) corresponding to $d = 0.035$ cm and 0.070 cm respectively ($b = 0.0088$ cm and $\lambda = 6.328 \times 10^{-5}$ cm).

or when,

$$d \sin \theta = 0, \lambda, 2\lambda, 3\lambda, \dots \quad (47)$$

The actual positions of the maxima will approximately occur at the above angles provided the variation of the diffraction term is not too rapid. Further, a maximum may not occur at all if θ corresponds to a diffraction minimum, i.e., if $b \sin \theta = \lambda, 2\lambda, 3\lambda, \dots$. These are usually referred to as missing orders. For example, in Fig. 16.27 we can see that for $b = 0.0088$ cm, the interference maxima are extremely weak around $\theta \approx 0.41^\circ$; this is because of the fact that at

$$\begin{aligned} \theta &= \sin^{-1} \left(\frac{\lambda}{b} \right) \\ &= \sin^{-1} \left[\frac{6.328 \times 10^{-5}}{8.8 \times 10^{-3}} \right] = \sin^{-1} [7.19 \times 10^{-3}] \\ &\approx 0.00719 \text{ radians} \\ &\approx 0.412^\circ \end{aligned}$$

the first minimum of the diffraction term occurs.

Example 16.9 Consider the case when $b = 8.8 \times 10^{-3}$ cm, $d = 7.0 \times 10^{-2}$ cm and $\lambda = 6.328 \times 10^{-5}$ cm (see Fig. 16.27). How many interference minima will occur between

the two diffraction minima on either side of the central maximum? In the experimental arrangement corresponding to Fig. 16.26 the screen was placed at a distance of $15'$. Calculate the fringe width.

Solution: The interference minima will occur when Eq. (46) is satisfied, i.e., when

$$\begin{aligned} \sin \theta &= \left(n + \frac{1}{2} \right) \frac{\lambda}{d} = 0.904 \times 10^{-3} \left(n + \frac{1}{2} \right); \\ n &= 0, 1, 2, \dots \\ &= 0.452 \times 10^{-3}, 1.356 \times 10^{-3}, 2.260 \times 10^{-3}, \\ &\quad 3.164 \times 10^{-3}, 4.068 \times 10^{-3}, 4.972 \times 10^{-3}, \\ &\quad 5.876 \times 10^{-3}, 6.780 \times 10^{-3} \end{aligned}$$

Thus there will be sixteen minima between the two first order diffraction minima.

The angular separation between two interference maxima is approximately given by (see Eq. 47)

$$\Delta \theta \approx \frac{\lambda}{d} = 0.904 \times 10^{-4}$$

Thus the fringe width will be

$$15 \times 12 \times 2.54 \times 0.904 \times 10^{-4} \approx 0.0413 \text{ cm}$$

16.7 N-SLIT FRAUNHOFER DIFFRACTION PATTERN

We next consider the diffraction pattern produced by N parallel slits, each of width b ; the distance between two consecutive slits is assumed to be d .

As before, we assume that each slit consists of n equally spaced point sources with spacing Δ (see Fig. 16.28). Thus the field at an arbitrary point P will essentially be a sum of N terms:

$$E = A \frac{\sin \beta}{\beta} \cos(\omega t - \beta) + A \frac{\sin \beta}{\beta} \cos(\omega t - \beta - \Phi_1) + \dots + A \frac{\sin \beta}{\beta} \cos(\omega t - \beta - (N-1)\Phi_1) \quad (48)$$

where the first term represents the amplitude produced by the first slit, the second term by the second slit, etc. and the various symbols have the same meaning as in Sec. 16.5. Rewriting Eq. (48) we get

$$E = \frac{A \sin \beta}{\beta} [\cos(\omega t - \beta) + \cos(\omega t - \beta + \Phi_1) + \dots + \cos(\omega t - \beta - (N-1)\Phi_1)] \\ = \frac{A \sin \beta}{\beta} \frac{\sin N\gamma}{\sin \gamma} \cos\left[\omega t - \beta - \frac{1}{2}(N-1)\Phi_1\right] \quad (49)$$

where

$$\gamma = \frac{\Phi_1}{2} = \frac{\pi}{\lambda} d \sin \theta$$

The corresponding intensity distribution will be

$$I = I_0 \frac{\sin^2 \beta}{\beta^2} \frac{\sin^2 N\gamma}{\sin^2 \gamma} \quad (50)$$

where $I_0 \sin^2 \beta / \beta^2$ represents the intensity distribution

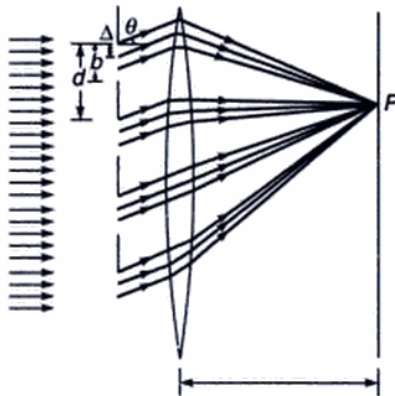


Fig. 16.28 Fraunhofer diffraction of a plane wave incident normally on a multiple slit.

produced by a single slit. As can be seen, the intensity distribution is a product of two terms; the first term $\left(\frac{\sin^2 \beta}{\beta^2}\right)$ represents the diffraction pattern produced by a single slit and the second term $\left(\frac{\sin^2 N\gamma}{\sin^2 \gamma}\right)$ represents the interference pattern produced by N equally spaced point sources. For $N = 1$, Eq. (50) reduces to the single slit diffraction pattern [see Eq. (10)] and for $N = 2$, to the double slit diffraction pattern [see Eq. (45)]. In Fig. 16.29 we have given a plot of the function

$$\frac{\sin^2 N\gamma}{\sin^2 \gamma}$$

as a function of γ for $N = 5$ and $N = 11$. One can immediately see that as the value of N becomes very large, the above function would become very sharply peaked at $\gamma = 0, \pi, 2\pi, \dots$. Between the two peaks, the function vanishes when

$$\gamma = \frac{p\pi}{N}; \quad p = \pm 1, \pm 2, \dots \quad \text{but} \quad p \neq 0, \pm N, \pm 2N$$

which are referred to as secondary minima.

16.7.1 Positions of Maxima and Minima

When the value of N is very large, one obtains intense maxima at $\gamma \approx m\pi$, i.e., when

$$d \sin \theta = m\lambda \quad (m = 0, 1, 2, \dots) \quad (51)$$

This can be easily seen by noting that

$$\lim_{\gamma \rightarrow m\pi} \frac{\sin N\gamma}{\sin \gamma} = \lim_{\gamma \rightarrow m\pi} \frac{N \cos N\gamma}{\cos \gamma} = \pm N;$$

thus, the resultant amplitude and the corresponding intensity distributions are given by

$$E = N \frac{A \sin \beta}{\beta} \quad (52)$$

and

$$I = N^2 I_0 \frac{\sin^2 \beta}{\beta^2} \quad (53)$$

where

$$\beta = \frac{\pi b \sin \theta}{\lambda} = \frac{\pi b}{\lambda} \frac{m\lambda}{d} = \frac{\pi b m}{d} \quad (54)$$

Such maxima are known as principal maxima. Physically, at these maxima the fields produced by each of the slits are in phase and, therefore, they add up and the resultant field is N times the field produced by each of the slits;

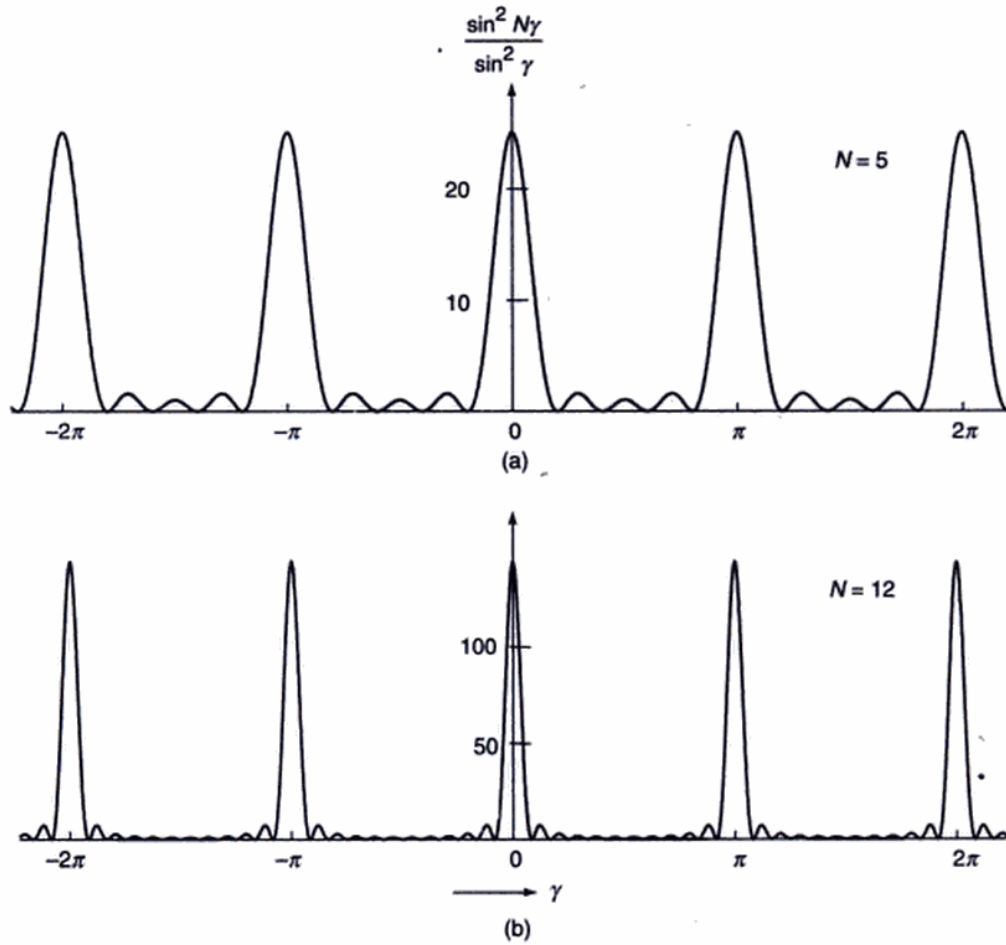


Fig. 16.29 The variation of the function $\sin^2(N\gamma) / \sin^2 \gamma$ with γ for $N = 5$ and 12 . As N becomes larger, the function would become more and more sharply peaked at $\gamma = 0, \pm \pi, \pm 2\pi, \pm 3\pi, \dots$

consequently, the intensity has a large value unless $\frac{\sin^2 \beta}{\beta^2}$ itself is very small. Since $|\sin \theta| \leq 1$, m cannot be greater than d/λ [see Eq. (51)]; thus, there will only be a finite number of principal maxima.

From Eq. (50) it can easily be seen that the intensity is zero when either

$$b \sin \theta = n\lambda, n = 1, 2, 3, \dots \quad (55)$$

or

$$N\gamma = p\pi, p \neq N, 2N, \dots \quad (56)$$

Equation (55) gives us the minima corresponding to the single slit diffraction pattern. The angles of diffraction corresponding to Eq. (56) are

$$d \sin \theta = \frac{\lambda}{N}, \frac{2\lambda}{N}, \dots, \frac{(N-1)\lambda}{N}, \frac{(N+1)\lambda}{N}, \frac{(N+2)\lambda}{N}, \dots, \frac{(2N-1)\lambda}{N}, \frac{(2N+1)\lambda}{N}, \frac{(2N+2)\lambda}{N}, \dots \quad (57)$$

Thus, between two principal maxima we have $(N-1)$ minima. Between two such consecutive minima the intensity has to have a maximum; these maxima are known as secondary maxima. Typical diffraction patterns for $N = 1, 2, 3$, and 4 are shown in Fig. 16.30 and the intensity distribution as predicted by Eq. (50) for $N = 4$ is shown in Fig. 16.31. When N is very large the principal maxima will be much more intense in comparison to the secondary maxima. We may mention here two points:

- (i) A particular principal maximum may be absent if it corresponds to the angle which also determines the minimum of the single-slit diffraction pattern. This will happen when

$$d \sin \theta = m\lambda \quad (58)$$

and

$$b \sin \theta = \lambda, 2\lambda, 3\lambda, \dots \quad (59)$$

are satisfied simultaneously and is usually referred to as a missing order. Even when Eq. (59) does not hold

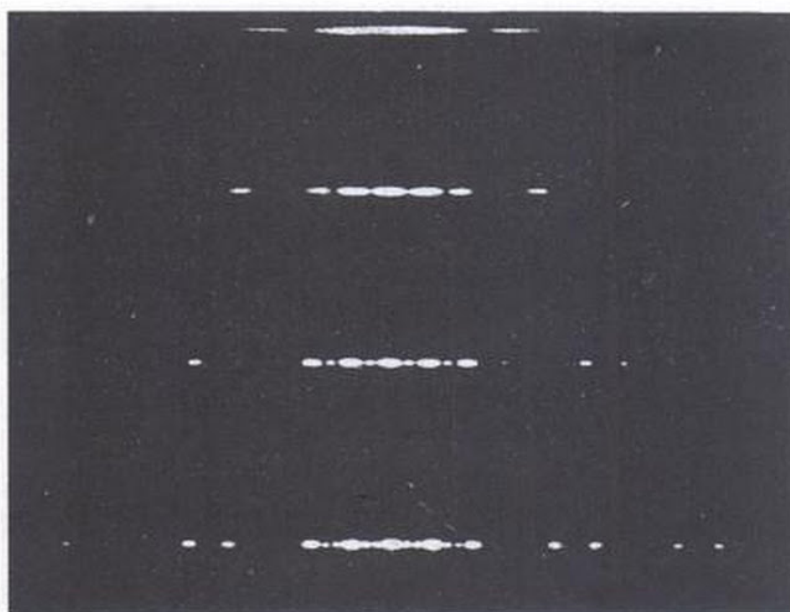


Fig. 16.30 The multiple-slit Fraunhofer diffraction patterns corresponding to $b = 0.0044$ cm, $d = 0.0132$ cm and $\lambda = 6.328 \times 10^{-5}$ cm. The number of slits are 1, 2, 3 and 4 respectively (After Ref. 17; used with permission).

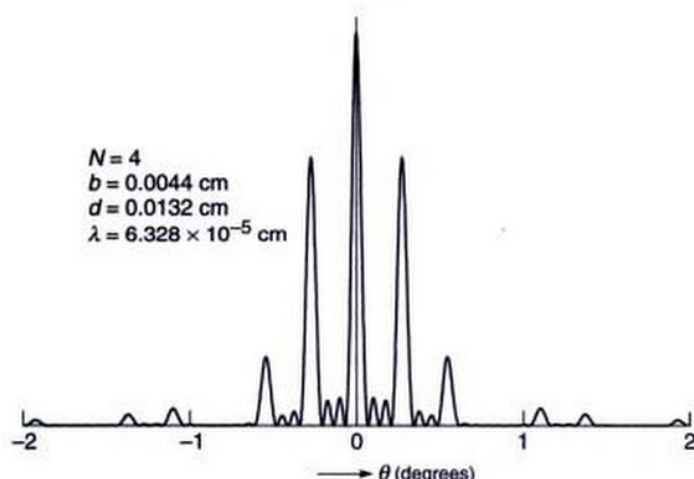


Fig. 16.31 The intensity distribution corresponding to the four-slit Fraunhofer diffraction pattern as predicted by Eq. (50) corresponding to $b = 0.0044$ cm, $d = 0.0132$ cm and $\lambda = 6.328 \times 10^{-5}$ cm. The principle maxima occur at $\theta \approx 0.275^\circ, 0.55^\circ, 0.82^\circ, 1.1^\circ, \dots$. Notice the (almost) absent third order.

exactly (i.e., if $b \sin \theta$ is close to an integral multiple of λ), the intensity of the corresponding principal maximum will be very weak (see, for example, Fig. 16.31 around $\theta \approx 0.8^\circ$).

- (ii) In addition to the minima predicted by Eq. (56), we will also have the diffraction minima (see Eq. 55);

however, when N is very large, the number of such minima will be very small.

16.7.2 Width of the Principal Maxima

We have shown above that in the diffraction pattern produced by N slits, the m th order principal maximum occurs at

$$d \sin \theta_m = m\lambda, \quad m = 0, 1, 2, \dots \quad (60)$$

Further, the minima occur at the angles given by Eq. (57). If $\theta_m + \Delta\theta_{1m}$ and $\theta_m - \Delta\theta_{2m}$ represent the angles of diffraction corresponding to the first minimum on either side of the principal maximum, then $\frac{1}{2}(\Delta\theta_{1m} + \Delta\theta_{2m})$ is known as the angular half width of the m th order principal maximum. For a large value of N , $\Delta\theta_{1m} \approx \Delta\theta_{2m}$ which we write as $\Delta\theta_m$. Clearly,

$$d \sin (\theta_m \pm \Delta\theta_m) = m\lambda \pm \frac{\lambda}{N} \quad (61)$$

But

$$\begin{aligned} \sin (\theta_m \pm \Delta\theta_m) &= \sin \theta_m \cos \Delta\theta_m \pm \cos \theta_m \sin \Delta\theta_m \\ &\approx \sin \theta_m \pm \Delta\theta_m \cos \theta_m \end{aligned} \quad (62)$$

Thus Eq. (61) gives us

$$\Delta\theta_m \approx \frac{\lambda}{Nd \cos \theta_m} \quad (63)$$

which shows that the principal maximum becomes sharper as N increases.

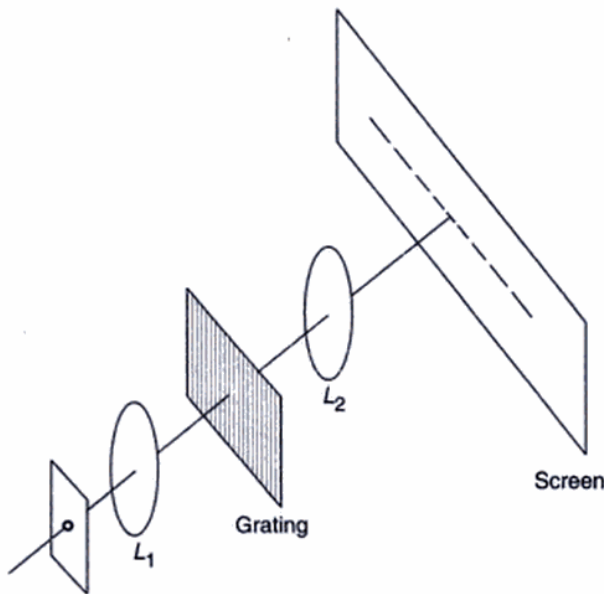


Fig. 16.32 Fraunhofer diffraction of a plane wave incident normally on a grating.

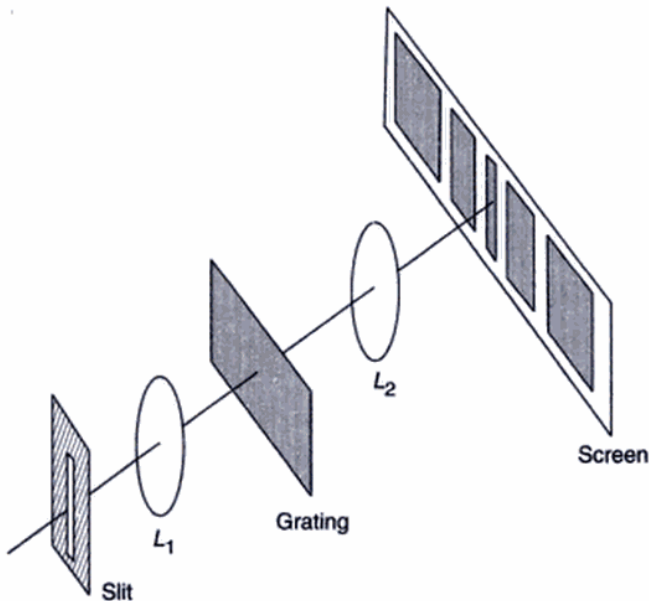


Fig. 16.33 If instead of a point source we have a slit in the focal plane of L_1 then one will obtain bands on the focal plane of L_2 .

source the second and third order spectra overlap. (ii) What will be the angular separation of the D_1 and D_2 lines of sodium in the second order spectra?

Solution: (i) The grating element is

$$d = \frac{2.54}{15000} = 1.69 \times 10^{-4} \text{ cm}$$

Let θ_{mv} and θ_{mr} represent the angles of diffraction for the m th order spectrum corresponding to the violet and red

colours respectively. Thus

$$\theta_{2v} = \sin^{-1} \frac{2 \times 4 \times 10^{-5}}{1.69 \times 10^{-4}} \approx \sin^{-1} 0.473 \approx 28.2^\circ$$

$$\theta_{2r} = \sin^{-1} \frac{2 \times 7 \times 10^{-5}}{1.69 \times 10^{-4}} \approx \sin^{-1} 0.828 \approx 55.90^\circ$$

and

$$\theta_{3v} = \sin^{-1} \frac{3 \times 4 \times 10^{-5}}{1.69 \times 10^{-4}} \approx \sin^{-1} 0.710 \approx 45.23^\circ$$

where we have assumed the wavelengths of the violet and red colours to be 4×10^{-5} cm and 7×10^{-5} cm respectively. Since $\theta_{2r} > \theta_{3v}$, the second and third order spectra will overlap. Further since $\sin \theta_{3r} > 1$, the third order spectrum for the red colour will not be observed.

(ii) Since $d \sin \theta = m\lambda$, we have for small $\Delta\lambda$:

$$(d \cos \theta) \Delta\theta = m(\Delta\lambda)$$

or

$$\begin{aligned} \Delta\theta &= \frac{m\Delta\lambda}{d \left\{ 1 - \left(\frac{m\lambda}{d} \right)^2 \right\}^{1/2}} \\ &\approx \frac{2 \times 6 \times 10^{-8}}{1.69 \times 10^{-4} \left[1 - \left(\frac{2 \times 6 \times 10^{-5}}{1.69 \times 10^{-4}} \right)^2 \right]^{1/2}} \\ &\approx 0.0010 \text{ radians} \approx 3.47' \end{aligned}$$

Thus, if we are using telescope of angular magnification 10, the two lines will appear to have an angular separation of $34.7'$.

16.8.2 Resolving Power of a Grating

In the case of a grating the resolving power refers to the power of distinguishing two nearby spectral lines and is defined by the following equation:

$$R = \frac{\lambda}{\Delta\lambda} \quad (66)$$

where $\Delta\lambda$ is the separation of two wavelengths which the grating can just resolve; the smaller the value of $\Delta\lambda$, the larger the resolving power.

The Rayleigh criterion (see Sec. 16.4) can again be used to define the limit of resolution. According to this criterion, if the principal maximum corresponding to the wavelength $\lambda + \Delta\lambda$ falls on the first minimum (on the either side of the

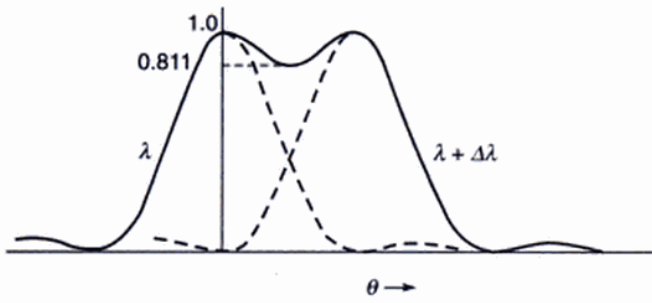


Fig. 16.34 The Rayleigh criterion for the resolution of two spectral lines.

principal maximum) of the wavelength λ , then the two wavelengths λ and $\lambda + \Delta\lambda$ are said to be just resolved (see Fig. 16.34). If this common diffraction angle is represented by θ and if we are looking at the m th order spectrum, then the two wavelengths λ and $\lambda + \Delta\lambda$ will be just resolved if the following two equations are simultaneously satisfied:

$$d \sin \theta = m(\lambda + \Delta\lambda) \quad (67)$$

and

$$d \sin \theta = m\lambda + \frac{\lambda}{N} \quad (68)$$

Thus

$$R = \frac{\lambda}{\Delta\lambda} = mN \quad (69)$$

which implies that the resolving power depends on the total number of lines in the grating—obviously on only those lines which are exposed to the incident beam (see the derivation in Sec. 16.6). Further, the resolving power is proportional to the order of the spectrum. Thus to resolve the D_1 and D_2 lines of sodium ($\Delta\lambda = 6 \text{ \AA}$) in the first order, N must be at least $(5.89 \times 10^{-5}) / (6 \times 10^{-8}) \approx 1,000$.

From Eq. (69) it appears that the resolving power of the grating would increase indefinitely if N is increased; however, for a given width of the grating $D (= Nd)$, as N is increased, d decreases and therefore the maximum value of m also decreases. Thus if d becomes 2.5λ , only first and second order spectra will be seen and if it is further reduced to about 1.5λ then only the first order spectrum will be seen.

16.8.3 Resolving Power of a Prism

We conclude this section by calculating the resolving power of a prism. Figure 16.35 gives a schematic description of the

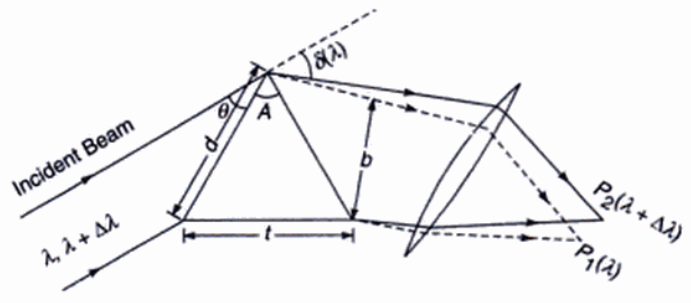


Fig. 16.35 The schematic of the experimental arrangement to observe the prism spectrum. P_1 and P_2 represent the images corresponding to λ and $\lambda + \Delta\lambda$ respectively.

experimental arrangement for observing the prism spectrum which is determined through the following formula:

$$n(\lambda) = \frac{\sin \frac{A + \delta(\lambda)}{2}}{\sin \frac{A}{2}} \quad (70)$$

where A represents the angle of the prism and δ the angle of minimum deviation. We assume that the refractive index decreases with λ (which is usually the case) so that δ also decreases with λ . In Fig. 16.35 the points P_1 and P_2 represent the images corresponding to λ and $\lambda + \Delta\lambda$ respectively. We are assuming that $\Delta\lambda$ is small so that the same position of the prism corresponds to the minimum deviation position for both wavelengths. In an actual experiment one usually has a slit source (perpendicular to the plane of the paper) forming line images at P_1 and P_2 . Since the faces of the prism are rectangular, the intensity distribution will be similar to that produced by a slit of width b (see Sec. 16.2)*. For the lines to be just resolved the first diffraction minimum [$m = 1$ in Eq. (12)] of λ should fall at the central maximum of $\lambda + \Delta\lambda$, thus we must have

$$\Delta\lambda \approx \frac{\lambda}{b} \quad (71)$$

In order to express $\Delta\delta$ in terms of $\Delta\lambda$, we differentiate Eq. (70):

$$\frac{dn}{d\lambda} = \frac{1}{\sin \frac{A}{2}} \cos \left[\frac{A + \delta(\lambda)}{2} \right] \frac{1}{2} \frac{d\delta}{d\lambda}$$

Thus

$$\Delta\delta = \frac{2 \sin \frac{A}{2}}{\cos \frac{A + \delta(\lambda)}{2}} \frac{dn}{d\lambda} \Delta\lambda$$

*Since we have a slit source we need not consider the diffraction in a direction perpendicular to the plane of the diagram.

Now from Fig. 16.35, we have

$$\theta = \frac{1}{2} [\pi - (A + \delta)]$$

or $\sin \theta = \frac{b}{a} = \cos \frac{A + \delta}{2}$

where the length a is shown in the figure. Further

$$\sin \frac{A}{2} = \frac{t/2}{a}$$

where t is the length of the base of the prism. Thus

$$\Delta \delta \approx \frac{t}{b} \frac{dn}{d\lambda} \Delta \lambda \quad (72)$$

Substituting in Eq. (72) we get for the resolving power

$$R = \frac{\lambda}{\Delta \lambda} = t \frac{dn}{d\lambda} \quad (73)$$

Now, for most glasses, the wavelength dependence of the refractive index (in the visible region of the spectrum) can be accurately described by the Cauchy formula

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} + \dots \quad (74)$$

Thus

$$\frac{dn}{d\lambda} = - \left[\frac{2B}{\lambda^3} + \frac{4C}{\lambda^5} + \dots \right] \quad (75)$$

the negative sign implying that the refractive index decreases with increase in wavelength. As an example we consider telescope crown glass for which*

$$A = 1.51375, B = 4.608 \times 10^{-11} \text{ cm}^2, C = 6.88 \times 10^{-22} \text{ cm}^4$$

For $\lambda = 6 \times 10^{-5} \text{ cm}$ we have

$$\begin{aligned} \frac{dn}{d\lambda} &\approx -[4.27 \times 10^2 + 3.54] \\ &\approx -4.30 \times 10^2 \text{ cm}^{-1} \end{aligned}$$

Thus, for $t = 2.5 \text{ cm}$ we have

$$R = \frac{\lambda}{\Delta \lambda} \approx 1000$$

which is an order of magnitude less than for typical diffraction gratings with 15,000 lines.

16.9 OBLIQUE INCIDENCE

Till now we have assumed plane waves incident normally on the grating. For experimental setting it is quite difficult to achieve the condition of normal incidence to a great precision and it is easily seen that slight deviations from normal incidence will introduce considerable errors. It is,

*Data quoted from Ref. 2.

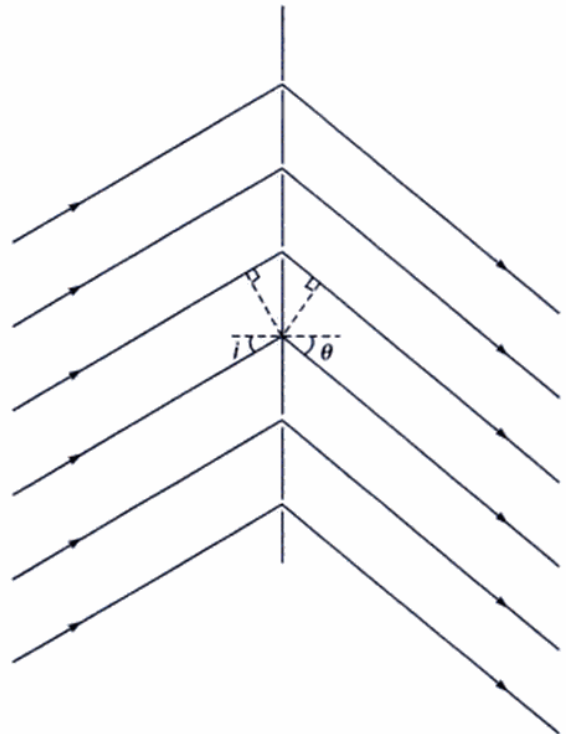


Fig. 16.36 Diffraction of a plane wave incident obliquely on a grating.

therefore, more practical to consider the more general oblique incidence case (see Fig. 16.36). The wavelength measurement can be carried out by using the method of minimum deviation as we do for prisms.

If the angle of incidence is i , then the path difference of the diffracted rays from two corresponding points in adjacent slits will be $d \sin \theta + d \sin i$ (see Fig. 16.36). Thus, principal maxima will occur when

$$d(\sin \theta + \sin i) = m\lambda \quad (76)$$

or $d[\sin (\delta - i) + \sin i] = m\lambda \quad (77)$

when $\delta = i + \theta$ is the angle of deviation. For δ to be minimum we must have

$$\frac{d}{di} [\sin (\delta - i) + \sin i] = 0 \quad (78)$$

$$-\cos (\delta - i) + \cos i = 0$$

i.e., $i = \delta - i = \theta \quad (79)$

or

$$i = \frac{\delta}{2} = \theta \quad (80)$$

Hence, at the position of minimum deviation, the grating condition becomes

$$2d \sin \frac{\delta}{2} = m\lambda \quad (81)$$

The minimum deviation position can be obtained in a manner similar to that used in the case of a prism and since the adjustments are relatively simpler, this provides a more accurate method for the determination of λ .

16.10 X-RAY DIFFRACTION*

Visible light is an electromagnetic wave whose wavelength approximately lies between 4000 Å and 7000 Å. X-rays are also electromagnetic waves whose wavelengths are ~ 1 Å. Obviously, it is extremely difficult to make slits which are narrow enough for the study of X-ray diffraction patterns. Since the interatomic spacings in a crystal are usually of the order of Angstroms, one can use it as a three-dimensional diffraction grating for studying the diffraction of X-rays. Indeed, X-rays have extensively been used to study crystal structures⁷.

In an ideal crystal, the atoms or molecules arrange themselves in a regular three-dimensional pattern which can be obtained by a three dimensional repetition of a certain unit pattern. This simplest volume which has all the characteristics of the whole crystal and which completely fills space is called the unit cell. One can think of various identifiable planes in the regular three-dimensional periodic arrangement. Miller indices are universally used as a system of notation for planes within a crystal. They specify the orientation of planes relative to the crystal axis without giving the position of the plane in space with respect to the origin. These indices are based on the intercepts of a plane with the three crystal axes, each intercept with an axis being measured in terms of unit cell dimensions (a , b or c) along that axis. To determine the Miller indices of a plane, the following procedure is used:

1. Find the intercepts (of the plane nearest to the origin) on the three axes and express them as multiple or fractions of the unit cell dimension.
2. Take the reciprocals of these numbers and multiply by the LCM of the denominators.
3. Enclose in parentheses.

For example, a (111) plane intercepts all three axes at one unit distance (see Fig. 16.37a); a (211) plane intercepts the three axes at $\frac{1}{2}$, 1 and 1 unit distances (see Fig. 16.37b). Similarly, a (110) plane intercepts the z -axis at ∞ . Miller indices can also be negative, the minus sign is shown above the digit like $(\bar{1}11)$. Figure 16.38 shows the planes characterized by the Miller indices $(\bar{1}11)$ in a simple cubic lattice.

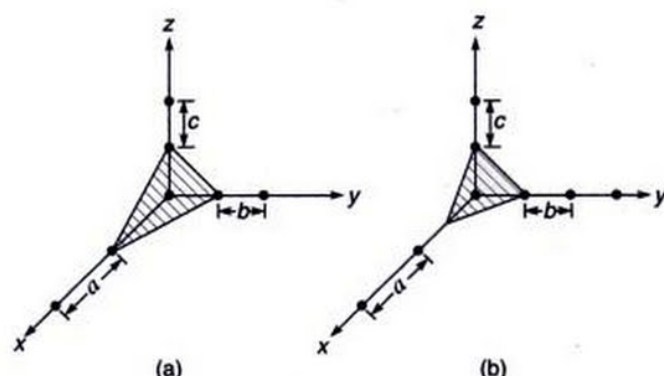


Fig. 16.37 (a) The (111) plane intercepts all three axes at one unit distance of each axial dimension. (b) The (211) plane intercepts the three axes at $\frac{1}{2}$, 1 and 1 unit distances.

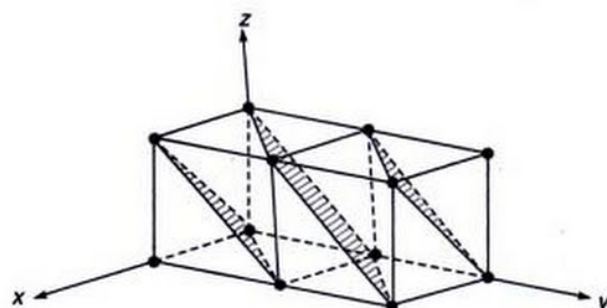


Fig. 16.38 Planes characterized by the Miller indices $(\bar{1}11)$ in a simple cubic lattice.

Consider a monochromatic beam of X-rays to be incident on a crystal. In Fig. 16.39 the horizontal dotted lines represent a set of parallel crystal planes with Miller indices (hkl) . $W_1 W_2$ and $W_3 W_4$ represent the incident and reflected wavefronts respectively. Obviously, the secondary wavelets emanating from the points A , B and C are in phase on $W_3 W_4$ (see Sec. 10.4 and Fig. 10.7); and the waves emanating from the points A_1 , B_1 and C_1 will also be in phase on $W_3 W_4$ if

$$XB_1 + B_1 Y = m\lambda, \quad m = 1, 2, 3, \dots \quad (82)$$

or when

$$2d_{hkl} \sin \theta = m\lambda \quad (83)$$

where d_{hkl} is the interplanar spacing between crystal planes of indices (hkl) , $m = 1, 2, 3, \dots$ is called the order of diffraction and θ is known as the glancing angle. This equation is known as Bragg's law and gives the angular positions of the reinforced diffracted beams in terms of the wavelength λ of the incoming X-rays and of the interplanar spacings d_{hkl} of the crystal planes. When the condition expressed by

*The author is grateful to Professor Lalit K. Malhotra for his help in writing this section.

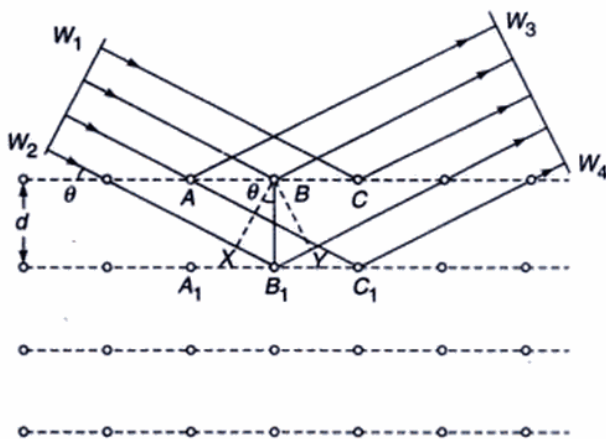


Fig. 16.39 Reflection of a plane wave by a set of parallel crystal planes characterized by the Miller indices (hkl) . When the Bragg condition $2d \sin \theta = m\lambda$ is satisfied, the waves scattered from different rows will be in phase.

Eq. (83) is not satisfied, destructive interference occurs and no reinforced beam will be produced. Constructive interference occurs when the condition given by Eq. (83) is satisfied leading to peaks in the intensity distribution. For solids which crystallize in cubic structures (which are discussed later), the interplanar spacing d_{hkl} between two closest parallel planes with Miller indices (hkl) is given by

$$d_{hkl} = \frac{a}{\sqrt{h^2 + k^2 + l^2}} \quad (84)$$

where a represents the lattice constant. Thus knowing the Miller indices, we can find d_{hkl} and from Bragg's law, we can determine the value of θ at which Bragg's equation can be satisfied.

There are three types of cubic structures: simple cubic, body centred cubic (BCC) and face centred cubic (FCC). Figure 16.39 shows a simple cubic structure (abbreviated as SC) in which the atoms are at the corners of a cube which forms what is known as a unit cell. The crystal is built up by the repetition of this unit cell in three dimensions. In addition, if there is an atom at the centre of each cube (shown as 9, 10, 11 and 12 in Fig. 16.40), the arrangement is known as a BCC structure. The distance between two adjacent planes characterized by the Miller indices $(\bar{1}10)$ is $a/\sqrt{2}$ which can be verified by simple geometry. On the other hand, if instead of having an atom at the center of the cube there is an atom at the center of each of the six faces of the cube (see Fig. 16.41) we will have the FCC structure. Copper, silver and gold crystallise in the FCC form with the

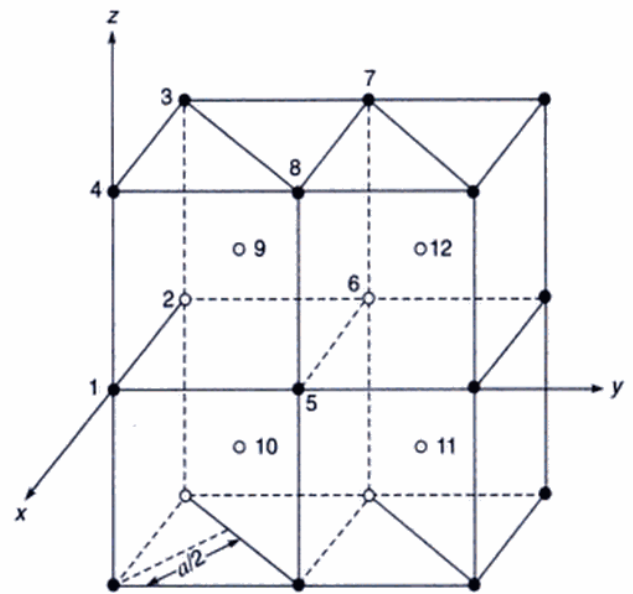


Fig. 16.40 A body centred cubic (bcc) lattice. The $(\bar{1}10)$ planes are separated by $a/\sqrt{2}$

lattice parameter $a = 3.61 \text{ \AA}$, 4.09 \AA and 4.08 \AA respectively. Metals like sodium, barium and tungsten crystallize in the BCC form with $a = 4.29 \text{ \AA}$, 5.03 \AA and 3.16 \AA respectively*.

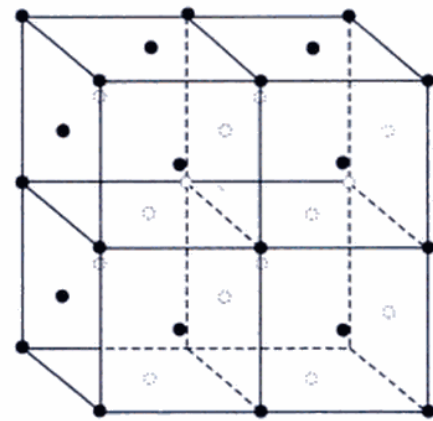


Fig. 16.41 A face centred cubic (fcc) lattice.

Just as there are optical missing orders of a diffraction grating, there are structural extinctions of X-ray reflection from a crystal. For simple cubic structures, reflections from all (hkl) planes are possible. However for the BCC structure, diffraction occurs only on planes whose Miller indices when added together total to an even number. Thus for the BCC structure, the principal diffracting planes for a first order diffraction are (110) , (200) , (211) (and other similar planes), etc. where $h + k + l$ is an even number. In the case

*Crystal structures other than cubic are also common; for example, zinc crystallizes into a hexagonal structure and carbon forms a diamond structure. However, the most important fact is that in all these structures there is a definite periodicity of atoms.

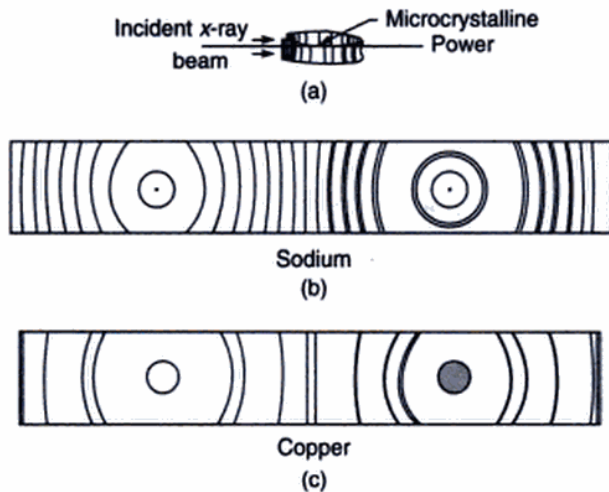


Fig. 16.43 (a) While using the powder method the photographic film is kept in a cylindrical form as shown in the figure. (b) and (c) represent schematic diffraction patterns for sodium and copper respectively.

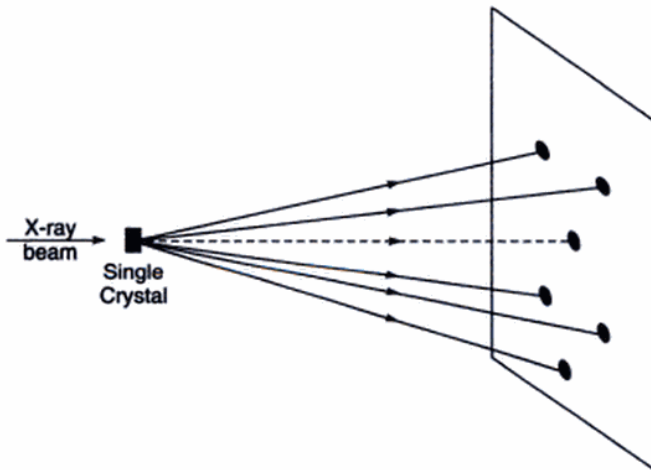


Fig. 16.44 When a polychromatic X-ray beam falls on a single crystal, one obtains Laue spots. Each set of planes chooses its own wavelength to satisfy the Bragg relation given by Eq. (84).

In order to calculate the angles of diffraction we substitute Eq. (84) in the Bragg's law [Eq. (83)] to obtain

$$\frac{2a}{\sqrt{h^2 + k^2 + l^2}} \sin \theta = m\lambda \quad (85)$$

We restrict ourselves only to first order reflections ($m = 1$); higher order reflections are usually rather weak (see also Problem 16.22). Thus Eq. (85) can be written in the form:

$$\sin \theta = \frac{\lambda}{2a} \sqrt{N} \quad (86)$$

*Based on Ref. 8; for a rigorous account, e.g., Ref. 9.

where

$$N = h^2 + k^2 + l^2$$

Now, for a simple cubic lattice, all values of (hkl) are possible implying the following possible values of N :

$$N = 1, 2, 3, 4, 5, 6, 7, \dots \text{ (SC)} \quad (87a)$$

Similarly, for a BCC lattice $h + k + l$ must be even implying

$$N = h^2 + k^2 + l^2 = 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, \dots \text{ (BCC)} \quad (87b)$$

Finally, for an FCC lattice, Miller indices are either all even or all odd implying

$$N = h^2 + k^2 + l^2 = 3, 4, 8, 11, 12, 16, 19, 20, 24, 27, \dots \text{ (FCC)} \quad (87b)$$

For a given structure and for given values of λ and a one can now easily calculate the different values of θ . For example, if we consider $\lambda = 1.540 \text{ \AA}$ and 1.544 \AA (corresponding to the $\text{CuK}_{\alpha 1}$ and $\text{CuK}_{\alpha 2}$ lines) then for sodium (which is a BCC structure with $a = 4.2906 \text{ \AA}$), the various values of θ are

(14.70°, 14.74°),	(21.03°, 21.09°),	(26.08°, 26.15°),
(30.50°, 30.59°),	(34.58°, 34.68°),	(38.44°, 38.56°),
(42.18°, 42.32°),	(45.88°, 46.03°),	(49.59°, 49.76°),
(53.38°, 53.58°),	(57.33°, 57.56°),	(61.54°, 61.82°),
(66.22°, 66.56°),	(79.41°, 80.23°),	

The two values inside the parentheses correspond to the two wavelengths 1.540 \AA and 1.544 \AA respectively. Because of the presence of two wavelengths one obtains double lines for each family of planes which become resolvable only at higher scattering angles. Similarly one can consider reflections from other structures (see Problems 16.19, 16.20 and 16.21). Each value of θ will give rise to a Debye-Scherrer ring shown in Figs. 16.42 (a), 16.43 (b) and 16.43 (c).

Finally, we should mention that the intensity of the diffracted wave depends on the number of atoms per unit area in the plane under consideration. For example, corresponding to the $(\bar{1}10)$ and the $(\bar{2}22)$ planes passing through a BCC lattice, there will be one atom and two atoms, respectively, in an area a^2 . Thus in the first case the intensity of the diffracted wave will be much more than in the second case.

16.11 THE SELF-FOCUSING PHENOMENON*

With the availability of intense laser beams, a large number of interesting non-linear optical phenomena have been investigated. One such non-linear phenomenon is the effect on the propagation of a light beam due to the dependence of

the refractive index on the intensity of the beam. This leads to the self-focusing (or defocusing) of the beam. In order to physically understand the self-focusing phenomenon we assume the non-linear dependence of the refractive index on the intensity to be of the form

$$n = n_0 + \frac{1}{2} n' E_0^2 \quad (88)$$

where n_0 is the refractive index of the medium in the absence of the electromagnetic field, n' is a constant representing the non-linear effect* and E_0 representing the amplitude of the electric field. As an example, we consider the incidence of a laser beam (propagating in the z -direction) having Gaussian intensity distribution in the transverse direction, i.e., we assume

$$E(x, y, z, t) \approx E_0 \cos(kz - \omega t) \quad (89)$$

with

$$E_0 = E_{00} \exp\left(-\frac{r^2}{a^2}\right) \quad (90)$$

where a represents the width of the Gaussian beam and $r (= \sqrt{x^2 + y^2})$ represents the cylindrical coordinate. In the absence of any nonlinear effects the beam will undergo diffraction divergence. However, if the beam is incident on a medium characterized by a positive value of n' , the intensity distribution will create a refractive index distribution which will have a maximum value on the axis (i.e., at $r=0$) and will gradually decrease with r . Indeed, using Eqs (88)–(90) we will have

$$\begin{aligned} n &\approx n_0 + \frac{1}{2} n' E_{00}^2 \exp\left(-\frac{2r^2}{a^2}\right) \\ &= \left(n_0 + \frac{1}{2} n' E_{00}^2\right) - \frac{1}{2} n_0 \left(\frac{r}{a}\right)^2 \end{aligned} \quad (91)$$

where

$$\alpha^2 = \frac{n_0 \alpha^2}{2n' E_{00}^2} \quad (92)$$

and in writing Eq. (91) we have expanded the exponential term and have retained only the first two terms. In other words, we are restricting ourselves to small values of r , which is the paraxial approximation. The term $\frac{1}{2} n' E_{00}^2$ is usually very small compared to n_0 ; so we may write (after squaring)

$$n^2 \approx n_0^2 \left[1 - \left(\frac{r}{a}\right)^2\right] \quad (93)$$

We may recall that in Sec. 2.4.1 we had considered propagation in a medium whose refractive index decreased parabolically from the axis and had shown that the beam could undergo periodic focusing (see Fig. 2.18). Indeed we had shown that the medium behaved like a converging lens of focal length $\pi a^2/2$ [see Eq. (48) of Chapter 2]. In the present case also because of nonlinear effects (with $n' > 0$), the medium will act as a converging lens of focal length approximately given by

$$f_{nl} = \frac{\pi}{2} \alpha \approx \frac{\pi}{2} \left[\frac{n_0}{2n' E_{00}^2} \right]^{1/2} a \quad (94)$$

the subscript (nl) signifying that the effect is due to a nonlinear phenomenon. Thus because of nonlinear effects the beam is said to undergo *self-focusing*; the word *self* signifies the fact that the beam creates its own refractive index gradient resulting in the focusing of the beam**.

Our analysis in Sec. 2.4.1 for the calculation of the focal length was based on ray optics and neglected diffraction effects. Now, in the absence of any non-linear effects, the beam will spread out due to diffraction and the angle of divergence will approximately be given by (see Fig. 16.13)

$$\theta_d \approx \frac{\lambda}{\pi a} = \frac{(\lambda_0/n_0)}{\pi a} \quad (95)$$

*This dependence may arise from a variety of mechanisms, such as the Kerr effect, electrostriction, thermal effect, etc. The simplest to understand is the thermal effect which is due to the fact that when an intense optical beam having a transverse distribution of intensity propagates through an absorbing medium, a temperature gradient is set up. For example, if the beam has a Gaussian transverse intensity variation (i.e., of the form $\exp(-r^2/a^2)$; the direction of propagation being along the z -axis), then the temperature will be maximum on the axis (i.e., $r=0$) and will decrease with increase in the value of r . If $dn/dT > 0$, the refractive index will be maximum on the axis and the beam will undergo focusing; on the other hand if $dn/dT < 0$, the beam will undergo defocusing (see, e.g., Ref. 9).

The Kerr effect arises due to the anisotropic polarizability of liquid molecules (like CS_2). An intense light wave will tend to orient the anisotropically polarized molecules such that the direction of maximum polarizability is along the direction of the electric vector; this changes the dielectric constant of the medium. On the other hand, electrostriction (which is important in solids) is the force which a nonuniform electric field exerts on a material medium; this force affects the density of the material, which in turn affects the refractive index. Thus, a beam having nonuniform intensity distribution along its wavefront will give rise to a refractive index variation leading to the focusing (or defocusing) of the beam. For a detailed discussion on electrostriction and Kerr effect, refer to Refs 9–11.

**It should be mentioned that if n' were a negative quantity, the refractive index would have increased as we move away from the axis and the beam would have undergone defocusing. For example, if the refractive index decreases with increase in temperature the beam may undergo what is known as thermal defocusing.

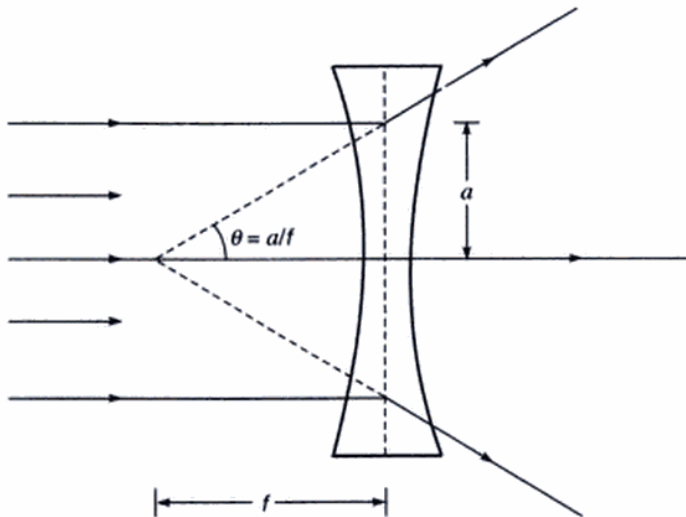


Fig. 16.45 When a plane wave is incident on a diverging lens, the transmitted rays diverge making an angle $\theta = a/f$ with the axis.

where λ_0 is the free space wavelength. Thus the phenomenon of diffraction can be approximated by a diverging lens of focal length (see Fig. 16.45).

$$f_d \approx \frac{a}{\theta_d} \approx \frac{1}{2} ka^2 \quad (96)$$

where $k = \frac{2\pi}{\lambda} = \frac{2\pi}{\lambda_0} n_0 \quad (97)$

Clearly if $f_d < f_{nl}$, the diffraction divergence will dominate and the beam will diverge. On the other hand, if $f_{nl} < f_d$, the non-linear focusing effects will dominate and the beam will undergo self-focusing. For $f_d \approx f_{nl}$, the two effects will cancel each other and the beam will propagate without any focusing or defocusing. This is the condition of *uniform waveguide like propagation*. In order to determine the critical power of the beam we note that the condition $f_d \approx f_{nl}$ implies

$$\frac{1}{2} ka^2 \approx \frac{\pi}{2} \left[\frac{n_0}{2n'E_{00}^2} \right]^{1/2} a$$

or

$$E_{00}^2 \approx \frac{1}{n_0 n'} \frac{\lambda_0^2}{8a^2} \quad (98)$$

Now the total power of the beam is given by

$$P = \int_0^\infty \text{velocity} \times (\text{energy/unit volume}) \times 2\pi r dr$$

$$\approx \int_0^\infty \left(\frac{c}{n_0} \right) \times \left(\frac{1}{2} \epsilon E_{00}^2 \right) \times 2\pi r dr$$

$$\approx \frac{c}{n_0} \left(\frac{1}{2} n_0^2 \epsilon_0 E_{00}^2 \right) \int_0^\infty \exp \left(-\frac{2r^2}{a^2} \right) 2\pi r dr$$

$$\approx \frac{\pi}{4} n_0 c \epsilon_0 E_{00}^2 a^2 \quad (99)$$

where $\epsilon (= n_0^2 \epsilon_0)$ is the dielectric permittivity of the medium and $\epsilon_0 (= 8.85 \times 10^{-12} \text{ C}^2/\text{N-m}^2)$ is the dielectric permittivity of free space (see Section 19.2). Substituting the expression for E_{00}^2 from Eq. (98) in Eq. (99) we obtain the following expression for the critical power:

$$P_{cr} \approx \frac{\pi}{32} (c \epsilon_0) \frac{\lambda_0^2}{n'} \quad (100)$$

Garmire, Chiao and Townes (1966)¹² carried out experiments on the self-focusing of a ruby laser beam ($\lambda_0 = 0.6943 \mu\text{m}$) in CS_2 and found that the critical power was $25 \pm 5 \text{ kW}$. Equation (100) gives us

$$P_{cr} \approx \frac{3.14}{32} \times 3 \times 10^8 \times 8.85 \times 10^{-12} \times \frac{(0.6943 \times 10^{-6})^2}{2 \times 10^{-20}}$$

$$\approx 6.3 \text{ kW} \quad (101)$$

where we have used the following parameters for CS_2 : $n_0 \approx 1.6276$, $n' \approx 1.8 \times 10^{-11} \text{ cgs units} \approx 2 \times 10^{-20} \text{ mks units}$. [The mks unit for n' is $(\text{meter/volt})^2$.] Although the result is wrong by a factor of about 4, one does obtain the correct order; this is indeed the case for all order-of-magnitude calculations. Thus

- (i) when $P < P_{cr}$, the beam will diverge due to diffraction.
- (ii) when $P = P_{cr}$, the beam will propagate without divergence or convergence. This is the condition for uniform waveguide propagation.
- (iii) when $P > P_{cr}$, we may extrapolate that the beam will undergo focusing, which is indeed borne out by more rigorous analysis. This is known as the self-focusing of the beam.

We may mention that a detailed study of the self-focusing phenomenon is of considerable importance in laser induced fusion experiments where there is a nonlinear interaction of the laser beam with the plasma.

16.12 SPATIAL FREQUENCY FILTERING

If $g(x, y)$ represents the field distribution on the front focal plane of a corrected lens (i.e., on the plane P_1 in Fig. 16.46), then on the back focal plane P_2 of the lens one obtains the Fourier transform of $f(x, y)$; the z -axis represents the optical axis of the lens. Thus, if $G(x, y)$ represents the field

image pattern. Since the spots are spaced closely, it represents a high frequency noise and the overall image has much smaller frequencies associated with it. Thus if we put a transparency similar to that shown in Fig. 16.47 (a) and allow only the low frequency components to pass through (as shown in Fig. 16.46), we will obtain, in the plane P_3 , an image which does not contain the unwanted high frequency noise (see Fig. 16.47).

The subject of spatial frequency filtering finds applications in many other areas like contrast improvement, character recognition, etc.

16.13 THE FOURIER TRANSFORMING PROPERTY OF A THIN LENS

In this section we will derive the Fourier transforming property of a thin lens. [viz., Eq.(102)]. We will use the results which will be derived in Sec. 17.5; thus we would like the reader to first go through Sections 17.4 and 17.5.

We will first show that the effect of a thin lens of focal length f is to multiply the incident field distribution by a

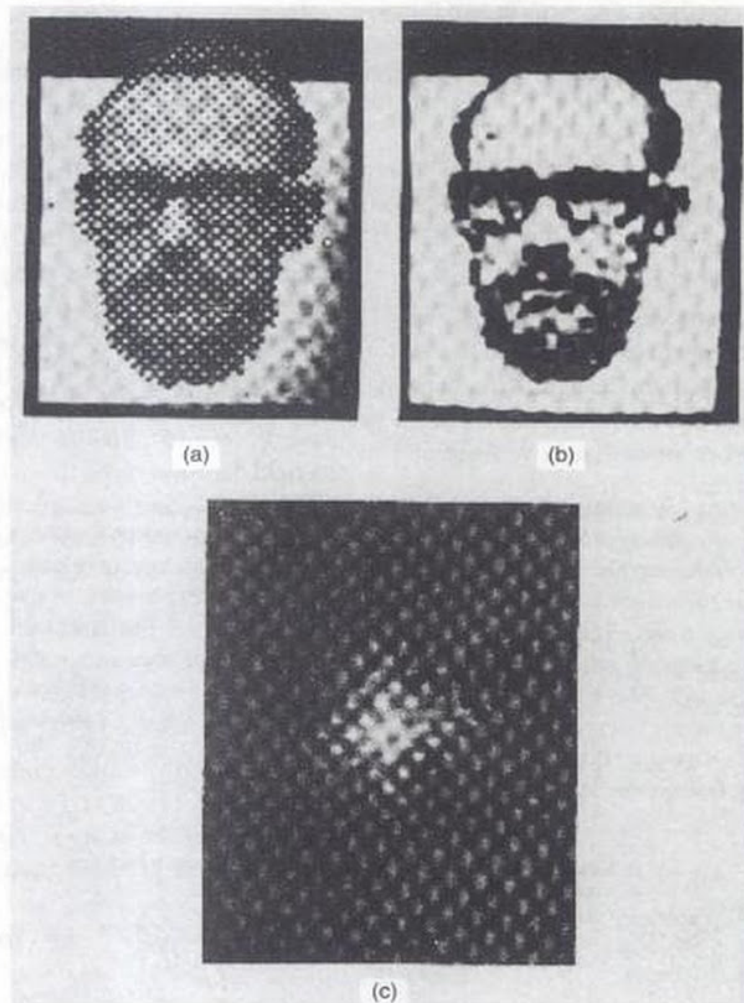


Fig. 16.47 (a) Shows a photograph consisting of regularly spaced black and white squares of varying sizes. When a pinhole is placed in the Fourier transform plane to block the high-frequency components, an image of the form shown in (c) is obtained; the frequency spectrum is shown in (b). Notice that in (c) shades of gray appear as well details such as the missing part of the eye glass frame (Adapted from Ref. 18).

factor p_L given by

$$p_L = \exp \left[-\frac{ik}{2f} (x^2 + y^2) \right] \quad (116)$$

Consider an object point O at a distance d_1 from an aberrationless thin lens of focal length f (see Fig. 16.48). If the image point I is at a distance d_2 from the lens, then d_2 is given by [see Sec. 3.4]:

$$\frac{1}{d_1} + \frac{1}{d_2} = \frac{1}{f} \quad (117)$$

where d_1 and d_2 represent the magnitude of the distances of the object and image points from the lens. The phase factor corresponding to the disturbance emanating from the point O is simply $\exp(+ikr)$, where r is the distance measured from the point O . Now

$$\begin{aligned} r &= (x^2 + y^2 + d_1^2)^{1/2} = d_1 \left[1 + \frac{x^2 + y^2}{d_1^2} \right]^{1/2} \\ &\approx d_1 + \frac{x^2 + y^2}{2d_1} \end{aligned}$$

wherein writing the last expression, we have assumed $x, y \ll d_1$, i.e., we have confined ourselves to a region close to the axis of the lens; this is known as the paraxial approximation. Thus, the phase distribution on the transverse plane P_2 at a distance d_1 from the point O (i.e., immediately in front of the lens—see Fig. 16.48) is given by

$$\exp(+ikr) \approx \exp \left[ik \left(d_1 + \frac{x^2 + y^2}{2d_1} \right) \right]$$

Since the image is formed at I , the incident spherical wave emerges as another spherical wave of radius d_2 , which under the paraxial approximation is

$$\exp \left[-ik \left(d_2 + \frac{x^2 + y^2}{2d_2} \right) \right]$$

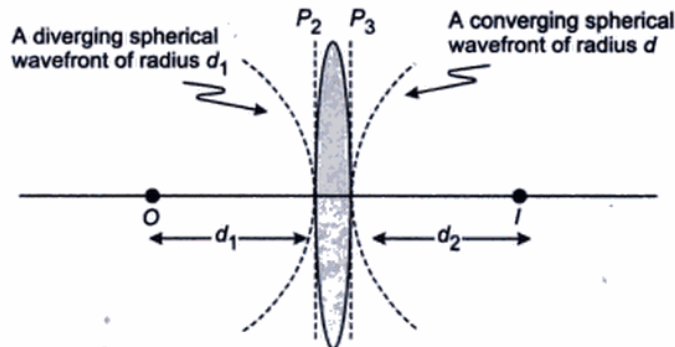


Fig. 16.48 Spherical waves emanating from an object point O , after refraction through a convex lens, emerge as spherical waves converging to the image point I

The negative sign in the exponent refers to the fact that we now have a converging spherical wave. Thus, if p_L represents the factor that when multiplied to the incident phase distribution gives the phase distribution of the emergent wave, then

$$\exp \left[-ik \left(d_2 + \frac{x^2 + y^2}{2d_2} \right) \right] = \exp \left[+ik \left(d_1 + \frac{x^2 + y^2}{2d_1} \right) \right] p_L$$

or

$$p_L = \exp[-ik(d_1 + d_2)] \exp \left[-\frac{ik}{2} \left(\frac{1}{d_1} + \frac{1}{d_2} \right) (x^2 + y^2) \right] \quad (118)$$

the subscript L on p corresponds to the fact that we are referring to a lens. If we use Eq. (117) and neglect the first factor in the above equation, because it is independent of x and y , we obtain Eq. (116). Thus the effect of a thin lens on an incident field is to multiply the incident phase distribution by a factor that is given by Eq. (116). For a plane wave incident along the axis, the emerging disturbance will simply be p_L , which can be seen to be the paraxial approximation of a converging spherical wavefront of radius f .

Now, let $g(x, y)$ represent the field distribution on the plane P_1 (see Fig. 16.49). We would first like to determine the field distribution on the plane P_2 i.e., at a distance f from the plane P_1 (see Fig. 16.49). Obviously the field will undergo Fresnel diffraction and on plane P_2 it will be given by [using Eq. (35) of Chapter 17]

$$\begin{aligned} u(x, y)|_{P_2} &= -\frac{i}{\lambda f} \exp(ikf) \iint g(\xi, \eta) \\ &\quad \times \exp \left\{ \frac{ik}{2f} [(x - \xi)^2 + (y - \eta)^2] \right\} d\xi d\eta \quad (119) \end{aligned}$$

Now, as shown earlier in this section, the effect of a thin lens of focal length f is to multiply the incident field distribution by the factor p_L given by Eq. (116), thus on the plane P_3 , the field distribution will be given by

$$\begin{aligned} u(x, y)|_{P_3} &= -\left(\frac{i}{\lambda f} \right) e^{ikf} \exp[-i\alpha(x^2 + y^2)] \iint g(\xi, \eta) \\ &\quad \times \exp[i\alpha[(x - \xi)^2 + (y - \eta)^2]] d\xi d\eta \quad (120) \end{aligned}$$

where

$$\alpha = \frac{k}{2f} = \frac{\pi}{\lambda f} \quad (121)$$

From plane P_3 the field will again undergo Fresnel diffraction and therefore on plane P_4 it will be given by [using Eq. (35) of Chapter 17]

$$\begin{aligned} u(x, y)|_{P_4} &= -\frac{i}{\lambda f} e^{ikf} \iint u(\zeta, \tau) \Big|_{P_3} \\ &\quad \exp[i\alpha[(x - \zeta)^2 + (y - \tau)^2]] d\zeta d\tau \quad (122) \end{aligned}$$

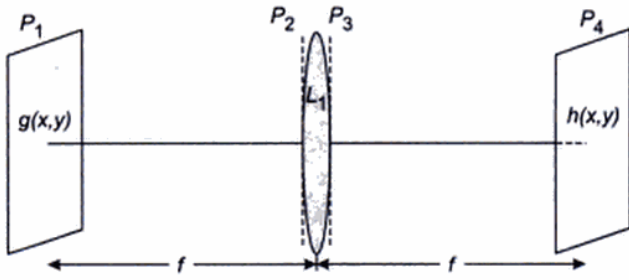


Fig. 16.49 A field distribution $g(x, y)$ placed at the front focal plane of a lens produces a field distribution $h(x, y)$ in the plane P_4 at the back focal plane of the lens. The field $g(x, y)$ first undergoes Fresnel diffraction from plane P_1 to P_2 , then it gets multiplied by a phase factor due to the presence of the lens, and the resultant field again undergoes Fresnel diffraction from plane P_3 to P_4 to produce the field distribution $h(x, y)$.

Substituting for $u|_{P_3}$ from Eq. (120) we get

$$u(x, y)|_{P_4} = \left[\left(-\frac{i}{\lambda f} \right) e^{ikf} \right]^2 I(x, y) \quad (123)$$

where

$$I(x, y) = \int \int_{-\infty}^{+\infty} g(\xi, \eta) H(x, y, \xi, \eta) d\xi d\eta \quad (124)$$

$$\begin{aligned} H(x, y, \xi, \eta) &= \int \int_{-\infty}^{+\infty} \exp \{ -i\alpha(\zeta^2 + \tau^2) \} \\ &\quad \times \exp \{ i\alpha[(\zeta - \xi)^2 + (\tau - \eta)^2] \} \\ &\quad \times \exp \{ i\alpha[x - \zeta]^2 + (y - \tau)^2 \} d\zeta d\tau \\ &= H_\xi(x) H_\eta(y) \end{aligned} \quad (125)$$

$$H_\xi(x) = \int_{-\infty}^{+\infty} \exp \{ i\alpha[\xi^2 - 2\xi\zeta + x^2 - 2x\zeta + \zeta^2] \} d\zeta$$

and a similar expression for H_η . Now,

$$\begin{aligned} \xi^2 - 2\xi\zeta + x^2 - 2x\zeta + \zeta^2 &= \zeta^2 - 2\zeta(x + \xi) + (x + \xi)^2 \\ &\quad - (x + \xi)^2 + \xi^2 + x^2 \\ &= (\zeta - g)^2 - 2x\xi \end{aligned}$$

where $g = x + \xi$. Thus

$$H_\xi = \exp[-2i\alpha x\xi] \int_{-\infty}^{+\infty} \exp[i\alpha(\zeta - g)^2] d\zeta$$

or

$$H_\xi(x) = e^{-2i\alpha x\xi} \sqrt{\frac{\pi}{-i\alpha}} \quad (126)$$

and a similar expression for $H_\eta(y)$. Thus

$$I(x, y) = \int \int_{-\infty}^{+\infty} g(\xi, \eta) H_\xi(x) H_\eta(y) d\xi d\eta$$

$$\begin{aligned} &= \frac{\pi}{-i\alpha} \int \int_{-\infty}^{+\infty} g(\xi, \eta) e^{-2i\alpha(x\xi + y\eta)} d\xi d\eta \\ &= i\lambda f \int \int_{-\infty}^{+\infty} g(\xi, \eta) e^{-i(u\xi + v\eta)} d\xi d\eta \end{aligned}$$

where we have used Eq. (121) and

$$u = 2\alpha x = \frac{2\pi x}{\lambda f} \quad \text{and} \quad v = 2\alpha y = \frac{2\pi y}{\lambda f} \quad (127)$$

represent the spatial frequencies in the x and y directions respectively. If we substitute the above expression for $I(x, y)$ in Eq. (123) we would obtain

$$u(x, y)|_{P_4} = \frac{1}{\lambda f} \int \int_{-\infty}^{+\infty} g(\xi, \eta) e^{-i(u\xi + v\eta)} d\xi d\eta \quad (128)$$

where we have neglected the unimportant constant phase factors. Equation (128) is the same as Eq. (103) gives the important result that

the field distribution on the back focal plane of a corrected lens is the Fourier transform of the field distribution on the front plane.

We should mention here that in writing the limits in the integral from $-\infty$ to $+\infty$ we have assumed the lens to be of infinite extent; the error involved is usually very small because in almost all practical cases

$$a/\lambda \gg 1 \quad (129)$$

where a represents the aperture of the lens.

Example 16.11 We consider a long narrow slit of width a (along the x -axis) placed on the front focal plane. Thus

$$\begin{aligned} g(x, y) &= A \quad |x| < \frac{a}{2} \\ &= 0 \quad |x| > \frac{a}{2} \end{aligned} \quad (130)$$

for all values of y ; see Fig. 16.3 (a). Substituting Eq. (130) in Eq. (128) we obtain

$$\begin{aligned} g(x, y) &= \frac{A}{\lambda f} \int_{-a/2}^{+a/2} e^{-iu\xi} d\xi \int_{-\infty}^{+\infty} e^{-iv\eta} d\eta \\ &= \frac{Aa}{\lambda f} \left(\frac{\sin \zeta}{\zeta} \right) 2\pi\delta(v) \end{aligned}$$

where

$$\zeta = \frac{ua}{2} = \frac{\pi ax}{\lambda f} = \left(\frac{\pi a}{\lambda} \right) \sin \theta$$

and θ is the angle of diffraction along the x -direction. We thus obtain the single slit diffraction pattern and the intensity is zero except on the x -axis [see Fig. 16.3(a)].

the lens. Repeat the calculations for a lens of same focal length but diameter 15 cm. Interpret the results physically.

[Ans. 1.46×10^{-4} cm, 4.88×10^{-5} cm]

- 16.7** Consider a set of two slits each of width $b = 5 \times 10^{-2}$ cm and separated by a distance $d = 0.1$ cm, illuminated by a monochromatic light of wavelength 6.328×10^{-5} cm. If a convex lens of focal length 10 cm is placed beyond the double slit arrangement, calculate the positions of the minima inside the first diffraction minimum.

[Ans. 0.0316 mm, 0.094 mm]

- 16.8** Show that when $b = d$, the resulting diffraction pattern corresponds to a slit of width $2b$.
- 16.9** Show that the first order and second order spectra will never overlap when the grating is used for studying a light beam containing wavelength components from 4000 Å to 7000 Å.
- 16.10** Consider a diffraction grating of width 5 cm with slits of width 0.0001 cm separated by a distance of 0.0002 cm. What is the corresponding grating element? How many order would be observable at $\lambda = 5.5 \times 10^{-5}$ cm? Calculate the width of the principal maximum. Would there be any missing orders?
- 16.11** For the diffraction grating of Problem 16.10, calculate the dispersion in the different orders. What will be the resolving power in each order?
- 16.12** A grating (with 15000 lines per inch) is illuminated by white light. Assuming that white light consists of wavelengths lying between 4000 and 7000 Å, calculate the angular widths of the first and the second order spectra. [Hint: You should not use Eq. (65); why?]
- 16.13** A grating (with 15,000 lines per inch) is illuminated by sodium light. The grating spectrum is observed on the focal plane of a convex lens of focal length 10 cm. Calculate the separation between the D_1 and D_2 lines of sodium. (The wavelengths of the D_1 and D_2 lines are 5890 and 5896 Å respectively.) [Hint: You may use Eq. (65).]
- 16.14** Calculate the resolving power in the second order spectrum of a 1 inch grating having 15,000 lines.
- 16.15** Consider a wire grating of width 1 cm having 1,000 wires. Calculate the angular width of the second order principal maxima and compare the value with the one corresponding to a grating having 5000 lines in 1 cm. Assume $\lambda = 5 \times 10^{-5}$ cm.
- 16.16** In the minimum deviation position of a diffraction grating the first order spectrum corresponds to an angular deviation of 30° . If $\lambda = 6 \times 10^{-5}$ cm, calculate the grating element.

- 16.17** Calculate the diameter of a telescope lens if a resolution of 0.1 seconds of arc is required at $\lambda = 6 \times 10^{-5}$ cm.
- 16.18** Assuming that the resolving power of the eye is determined by diffraction effects only, calculate the maximum distance at which two objects separated by a distance of 2 m can be resolved by the eye. (Assume pupil diameter to be 2 mm and $\lambda = 6000$ Å.)
- 16.19** (a) A pinhole camera is essentially a rectangular box with a tiny pinhole in front. An inverted image of the object is formed on the rear of the box. Consider a parallel beam of light incident normally on the pinhole. If we neglect diffraction effects then the diameter of the image will increase linearly with the diameter of the pinhole. On the other hand, if we assume Fraunhofer diffraction, then the diameter of the first dark ring will go on increasing as we reduce the diameter of the pinhole. Find the pinhole diameter for which the diameter of the geometrical image is approximately equal to the diameter of the first dark ring in the Airy pattern. Assume $\lambda = 6000$ Å and a separation of 15 cm between the pinhole and the rear of the box. (b) Figure 16.48 shows the quality of the image formed for various values of the diameter of the pinhole. Discuss qualitatively the fact that the image will get blurred if the diameter of the pinhole is too big or too small.

[Ans. (a) 0.47 mm]



Fig. 16.50 The image formed in a pinhole camera for different diameters of the pinhole. [Photograph downloaded from the internet by Professor K Thyagarajan; Ref: <http://www.cs.berkeley.edu/~daf/book/chapter-4.pdf>].

- 16.20** Copper is an FCC structure with lattice constant 3.615 Å. An X-ray powder photograph of copper is taken. The X-ray beam consists of wavelengths 1.540 Å and 1.544 Å. Show that diffraction maxima

Chapter 17

Fresnel Diffraction

One of your commissioners, M. Poisson, had deduced from the integrals reported by the author [Fresnel] the singular result that the centre of the shadow of an opaque circular screen must, when the rays penetrate there at incidences which are only a little oblique, be just as illuminated as if the screen did not exist. The consequence have been submitted to the test of a direct experiment, and observation has perfectly confirmed the calculation.

— Dominique Arago to the French Academy of Sciences*

Important Milestones

- | | |
|------|--|
| 1816 | Augustin Fresnel developed the theory of diffraction using the wave theory of light. |
| 1817 | Using Fresnel's theory, Poisson predicted a bright spot at the center of the shadow of an opaque disc—this is usually referred to as the 'Poisson spot'. |
| 1818 | Fresnel and Arago carried out the experiment to demonstrate the existence of the Poisson spot validating the wave theory. |
| 1874 | Marie Cornu developed a graphical approach to study Fresnel diffraction—this came to be known as the Cornu's spiral. |

17.1 INTRODUCTION

In the previous chapter we had mentioned that the phenomenon of diffraction can be broadly classified under two categories: under the first category comes the Fresnel class of diffraction in which either the source or the screen (or both) are at a finite distance from the diffracting aperture. In the second category comes the Fraunhofer class of diffraction (discussed in the previous chapter) in which the wave incident on the aperture is a plane wave and the diffraction pattern is observed on the focal plane of a convex lens, so that the screen is effectively at an infinite distance from the aperture. In the present chapter we will discuss the Fresnel class of diffraction and also study the transition to the Fraunhofer region. The underlying principle in the entire analysis is the Huygens–Fresnel principle according to which:

Each point on a wavefront is a source of secondary disturbance and the secondary wavelets emanating from different points mutually interfere.

In order to appreciate the implications of this principle we consider the incidence of a plane wave on a circular hole

*The author found this quotation in Ref. 1.

of radius a as shown in Fig. 17.1. In Sec. 16.3 we had shown that the beam will undergo diffraction divergence and the angular spreading will be given by

$$\Delta\theta \sim \frac{\lambda}{2a}$$

Thus, when $a \gg \lambda$ the intensity at a point R (which is deep inside the geometrical shadow) will be negligible; on the other hand, if $a \sim \lambda$ there will be almost uniform spreading

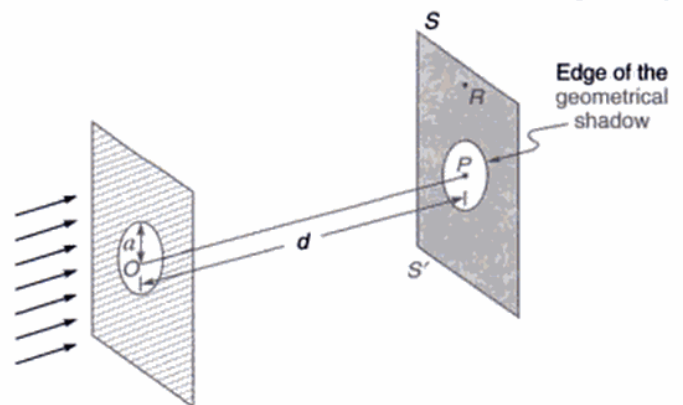


Fig. 17.1 Diffraction of a plane wave incident normally on a circular aperture of radius a .

out of the beam resulting in an (almost) uniform illumination of the screen. This phenomenon is a manifestation of the fact that when $a \gg \lambda$, the secondary wavelets emanating from different points on the circular aperture so beautifully interfere to produce (almost) zero intensity in the geometrical shadow and a large intensity inside the circular region (see Fig. 17.1). However, if $a \sim \lambda$ then the aperture almost acts as a point source resulting in a uniform illumination of the screen (see Fig. 10.3).

We will first introduce the concept of Fresnel half-period zones to have a qualitative understanding of the Fresnel diffraction pattern; this will be followed by a more rigorous analysis of the Fresnel class of diffraction and its transition to the Fraunhofer region.

17.2 FRESNEL HALF-PERIOD ZONES

Let us consider a plane wavefront WW' propagating in the z -direction as shown in Fig. 17.2. In order to determine the field at an arbitrary point P due to the disturbances reaching from different portions of the wavefront, we make the following construction: from the point P we drop a perpendicular PO on the wavefront. If $PO = d$, then with point P as centre we draw spheres of radii $d + \lambda/2$, $d + 2\lambda/2$, $d + 3\lambda/2$, ...; these spheres will intersect WW' in circles as shown in Fig. 17.2. The radius of the n th circle will obviously be given by

$$r_n = \left[\left(d + n \frac{\lambda}{2} \right)^2 - d^2 \right]^{1/2}$$

$$= \sqrt{n \lambda d} \left[1 + \frac{n \lambda}{4d} \right]^{1/2}$$

or

$$r_n \approx \sqrt{n \lambda d} \quad (1)$$

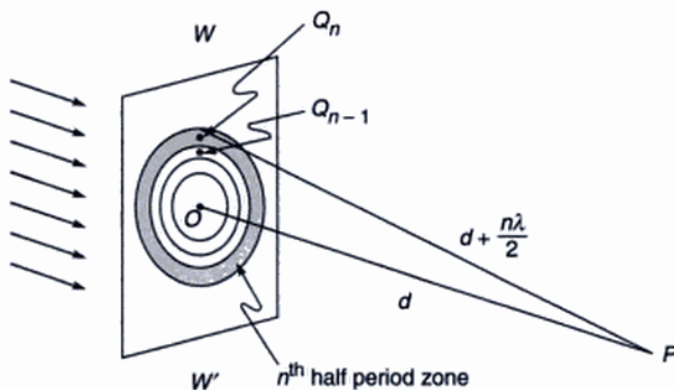


Fig. 17.2 Construction of Fresnel half-period zones.

* See, e.g., Ref. 2.

where we have assumed $d \gg \lambda$, this is indeed justified for practical systems using visible light; of course, we are assuming that n is not a very large number. The annular region between the n th circle and $(n-1)$ th circle is known as the n th half period zone; the area of the n th half-period zone is given by

$$A_n = \pi r_n^2 - \pi r_{n-1}^2$$

$$\approx \pi [n \lambda d - (n-1) \lambda d] = \pi \lambda d \quad (2)$$

Thus the areas of all the half-period zones are approximately equal. Now the resultant disturbance produced by the n th zone will be π out-of-phase with the disturbance produced by the $(n-1)$ th [or the $(n+1)$ th] zone. This can easily be seen from the following consideration: For infinitesimal area surrounding a point Q_n in the n th half period zone, there is a corresponding infinitesimal area surrounding the point Q_{n-1} in the $(n-1)$ th half-period zone such that

$$Q_n P - Q_{n-1} P = \frac{\lambda}{2}$$

which corresponds to a phase difference of π . Since the areas of the zones are approximately equal, one can have a one-to-one correspondence between points in various zones. Thus, the resultant amplitude at the point P can be written as

$$u(P) = u_1 - u_2 + u_3 - u_4 + \dots + (-1)^{m+1} u_m + \dots \quad (3)$$

where u_n represents the net amplitude produced by the secondary wavelets emanating from the n th zone; the alternate negative and positive signs represent the fact that the resultant disturbances produced by two consecutive zones are π out-of-phase with respect to each other. The amplitude produced by a particular zone is proportional to the area of the zone and inversely proportional to the distance of the zone from the point P ; further, it also depends on an obliquity factor which is proportional to $\frac{1}{2} (1 + \cos \chi)$ where χ is the angle that the normal to the zone makes with the line QP ; this obliquity factor comes out automatically from rigorous diffraction theory.* Thus we may write

$$u_n = \text{constant} \frac{A_n}{Q_n P} \frac{(1 + \cos \chi)}{2} \quad (4)$$

where A_n represents the area of the n th zone. It can be shown that if we use the exact expression for r_n , the area of the zones increase with n ; however, this slight increase in the area is exactly compensated by the increased distance of the zone from the point P . In spite of this, the amplitudes u_1, u_2, u_3, \dots decrease monotonically because of increased obliquity. Thus we may write

$$u_1 > u_2 > u_3 \dots \quad (5)$$

As a corollary of the above analysis we can consider a circular aperture of a fixed radius a and study the intensity variation along the axis. Whenever the distance

$$d = \frac{a^2}{(2n+1)\lambda}; n = 0, 1, 2, \dots \text{ (maxima)}$$

the point P (see Fig. 17.1) will correspond to a maximum. Similarly, when

$$d = \frac{a^2}{2n\lambda}; n = 1, 2, \dots \text{ (minima)}$$

the point P will correspond to a minimum. The intensity distribution on a screen SS' at off-axis points can be approximately calculated by using the half-period zones, but such a calculation is fairly cumbersome. However, from the symmetry of the problem, one can deduce that the diffraction pattern has to be in the form of concentric circular rings with their centres at the point P .

17.2.2 Diffraction by an Opaque Disc – The Poisson Spot

If instead of the circular aperture we have a circular disc [see Fig. 17.3(a)] and if the disc obstructs the first p half-period zones then the field at the point P would be

$$\begin{aligned} u(P) &= u_{p+1} - u_{p+2} + \dots \\ &= \frac{u_{p+1}}{2} \end{aligned} \quad (13)$$

Thus, we should always obtain a bright spot on the axis behind a circular disc; (the more rigorous theory also predicts the same result — see Sec. 17.4.2). This is called

the ‘Poisson spot’. We may mention here that it was in 1816 that the French physicist Augustin Fresnel developed the mathematical theory of diffraction using the wave theory of light. Simeon Poisson, the famous mathematician, used Fresnel's theory to predict a bright spot at the center of the shadow of an opaque disc. Poisson was a great supporter of the corpuscular theory of light and he said that since the bright spot is against common sense, the wave theory must be wrong. Shortly afterwards, Fresnel and Arago carried out the experiment to demonstrate the existence of the Poisson spot [see Fig. 17.3(b)], validating the wave theory.

17.3 THE ZONE-PLATE

A beautiful application of the concept of Fresnel half-period zones lies in the construction of the zone-plate which consists of a large number of concentric circles whose radii are proportional to the square root of natural numbers and the alternate annular regions of which are blackened (see Fig. 17.4). Let the radii of the circles be $\sqrt{1}K, \sqrt{2}K, \sqrt{3}K, \sqrt{4}K, \dots$ where K is a constant and has the dimension of length. We consider a point P_1 which is at a distance K^2/λ from the zone plate; for this point the blackened rings correspond to the 2nd, 4th, 6th, ... half-period zones. Thus, the even zones are obstructed and the resultant amplitude at P_1 [see Fig. 17.5 (a)] will be

$$u_1 + u_3 + u_5 + \dots \quad (14)$$

producing an intense maximum. For the point P_3 (which is at a distance $K^2/3\lambda$) the first blackened ring contains the 4th, 5th, 6th zones, the second blackened ring contains the 10th,

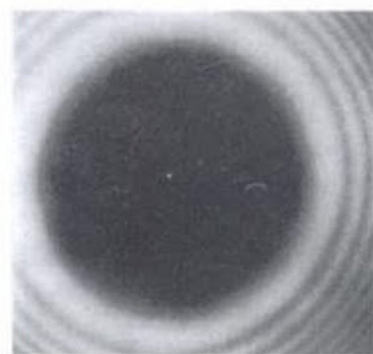
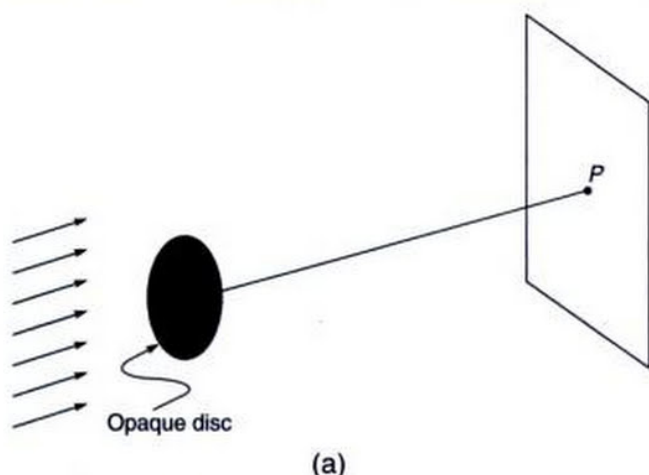


Fig. 17.3 (a) When a plane wave is incident normally on an opaque disc, a bright spot is always formed on an axial point. This spot is known as the Poisson spot. (b) The Poisson spot at the center of the shadow of a one cent coin; the screen is 20 m from the coin and the source of light is also 20 m from the coin [photograph adapted from Ref. 2]

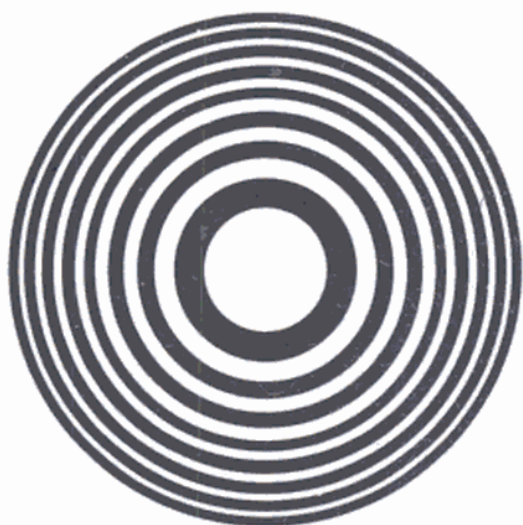


Fig. 17.4 The zone plate.

11th and 12th zones, etc.; thus the resultant amplitude would be

$$(u_1 - u_2 + u_3) + (u_7 - u_8 + u_9) + \dots \quad (15)$$

which would again correspond to a maximum, but it would not be as intense as the point P_1 . Between the points P_1 and P_3 there will be a point P_2 (at a distance $K^2/2\lambda$) where the resultant amplitude would be

$$(u_1 - u_2) + (u_5 - u_6) + \dots \quad (16)$$

implying that corresponding to P_2 the first blackened ring contains the 3rd and 4th half-period zones, etc. Obviously, the point P_2 will correspond to a minimum. Thus, if a plane wave is incident normally on a zone-plate, then the corresponding focal points are at distances

$$\frac{K^2}{\lambda}, \frac{K^2}{3\lambda}, \frac{K^2}{5\lambda}, \dots \quad (17)$$

from the zone-plate. Elementary calculations will show that the zone-plate suffers from considerable chromatic aberrations [see Problem 17.5].

Example 17.1 Assume a plane wave ($\lambda = 5 \times 10^{-5}$ cm) to be incident on a circular aperture of radius 0.5 mm. We will calculate the positions of the brightest and darkest points on the axis. For the brightest point, the aperture should contain only the first zone and thus we must have (see Fig. 17.1)

$$(0.05)^2 = OP \ (5 \times 10^{-5})$$

Thus $OP = 50$ cm. Similarly the darkest point would be at a distance

$$\frac{(0.05)^2}{2 \times 5 \times 10^{-5}} = 25 \text{ cm}$$

Example 17.2 Consider a zone-plate with radii

$$r_n = 0.1 \sqrt{n} \text{ cm}$$

For $\lambda = 5 \times 10^{-5}$ cm, we will calculate the positions of various foci. The most intense focal point will be at a distance

$$\frac{r_1^2}{\lambda} = \frac{0.01}{5 \times 10^{-5}} = 200 \text{ cm}$$

The other focal points will be at distances of 200/3, 200/5, 200/7 cm, etc. Between any two consecutive foci there will be dark points on the axis corresponding to which the first circle will contain an even number of half-period zones.

The zone-plate can also be used for imaging points on the axis, e.g., if we have a point source at S then a bright image will be formed at P , where the point P should be such that (see Fig. 17.5(b)):

$$SL + LP - SP = \frac{\lambda}{2} \quad (18)$$

the point L being on the periphery of the first circle of the zone plate [see Fig. 17.5(b)]. If the radius of the first circle is r_1 , then

$$\begin{aligned} SL + LP - SP &= \sqrt{a^2 + r_1^2} + \sqrt{b^2 + r_1^2} - (a + b) \\ &\approx a \left[1 + \frac{r_1^2}{2a^2} \right] + b \left[1 + \frac{r_1^2}{2b^2} \right] - (a + b) \\ &\approx \frac{r_1^2}{2} \left(\frac{1}{a} + \frac{1}{b} \right) \end{aligned} \quad (19)$$

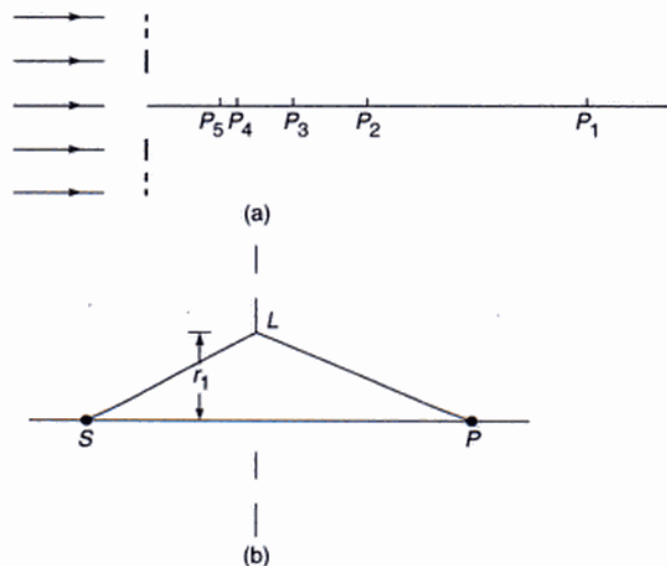


Fig. 17.5 (a) For a plane wave incident on a zone-plate, the maximum intensity occurs at the points P_1, P_3 etc; the minima occur at P_2, P_4, \dots (b) Imaging of a point object by a zone-plate.

Now

$$\rho^2 + d^2 = r^2$$

Thus

$$\rho d\rho = r dr$$

and Eq.(24) becomes

$$u(P) = -\frac{iA}{\lambda} \int_0^{2\pi} \int_d^{\sqrt{a^2+d^2}} e^{ikr} dr d\phi \quad (25)$$

The integration is very simple and since $k = 2\pi/\lambda$, we readily obtain

$$u(P) = A e^{ikd} (1 - e^{ip\pi}) \quad (26)$$

where we have defined p by the following equation

$$k [\sqrt{a^2 + d^2} - d] = p\pi$$

The above equation implies

$$QP - OP = \frac{p\lambda}{2}$$

where Q is a point on the periphery of the circular aperture (see Fig. 17.7). From Eq.(26) we readily get

$$I(P) = 4 I_0 \sin^2 \frac{p\pi}{2} \quad (27)$$

where I_0 is the intensity associated with the incident plane wave. Equation (27) tells us that the intensity is zero or maximum when p is an even or odd integer, i.e., when $QP - OP$ is an even or odd multiple of $\lambda/2$. This can be understood physically by using the concept of Fresnel half-period zones discussed in Sec. 17.2. Thus, if the aperture contains an even number of half-period zones, the intensity at the point P will be negligibly small and conversely, if the circular aperture contains an odd number of zones, the intensity at the point P will be maximum. Now, when $d \ll a$ (as is usually the case)

$$p \approx \frac{k}{\pi} \left[d \left(1 + \frac{a^2}{2d^2} \right) - d \right]$$

or

$$p \approx \frac{a^2}{\lambda d} \quad (28)$$

which is known as the Fresnel number of the aperture. In Fig. 17.8 we have plotted the corresponding intensity variation as a function of the dimensionless parameter

$$\frac{\lambda d}{a^2}$$

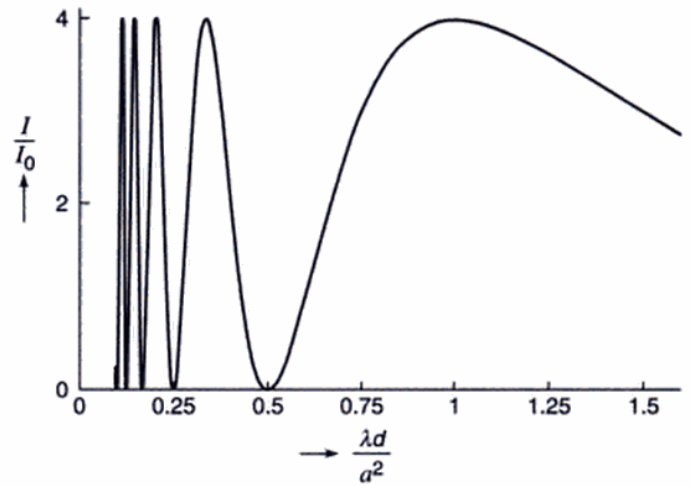


Fig. 17.8 The intensity variation on an axial point corresponding to a plane wave incident on a circular aperture of radius a .

The figure shows that when the (circular) aperture contains an even number of half-period zones, the intensity at the point P will be zero and when the aperture contains an odd number of zones, the intensity at the point P will be maximum.

17.4.2 Diffraction by a Circular Disc

We next consider the diffraction pattern produced by an opaque disc of radius a (see Fig. 17.3). Once again we will assume that the observation point lie on the axis of the disc. Equation (23) tells us that in order to calculate the field we have to carry out an integration over the open region of the aperture. Obviously, if $u_1(P)$ and $u_2(P)$ respectively represent the fields at the point P due to a circular aperture and an opaque disc (of the same radius), then

$$u_1(P) + u_2(P) = u_0(P) \quad (29)$$

where $u_0(P)$ represents the field in the absence of any aperture; Eq. (29) is known as the Babinet's principle. Thus,

$$\begin{aligned} u_2(P) &= u_0(P) - u_1(P) \\ &= u_0(P) - u_0(P)[1 - e^{ip\pi}] \end{aligned}$$

$$\text{or} \quad u_2(P) = u_0(P) e^{ip\pi} \quad (30)$$

where, for $u_1(P)$ we have used Eq.(26). Thus the intensity at the point P on the axis of a circular disc would be

$$I_2(P) = |u_2(P)|^2 = I_0(P) \quad (31)$$

which gives us the remarkable result that the intensity at a point on the axis of an opaque disc is equal to the intensity at the point in the absence of the disc! This is the Poisson spot discussed in Sec.17.2.2.

17.5 GAUSSIAN BEAM PROPAGATION

When a laser oscillates in its fundamental transverse mode, the transverse amplitude distribution is Gaussian. Also, the output of a single mode fiber is very nearly Gaussian. Therefore, the study of the diffraction of a Gaussian beam is of great importance. In Sec. 17.4 we have shown that for a plane wave incident normally on an aperture, the diffraction pattern is given by [see Eq. (23)]

$$u(P) = -\frac{iA}{\lambda} \iint \frac{1}{r} e^{ikr} d\xi d\eta \quad (32)$$

where the integration is over the area of the aperture (see Fig. 17.6) and

$$r = [(x - \xi)^2 + (y - \eta)^2 + z^2]^{1/2}$$

We assume the plane of the aperture to be $z = 0$. Now, if the amplitude and phase distribution on the plane $z = 0$ is given by $A(\xi, \eta)$ then the above integral is modified to

$$u(P) = -\frac{i}{\lambda} \iint A(\xi, \eta) \frac{e^{ikr}}{r} d\xi d\eta \quad (33)$$

The quantity r represents the distance between the point (ξ, η) on the plane of the aperture and the point P (whose coordinates are x, y, z) on the screen as shown in Fig. 17.6; obviously, the plane of the screen corresponds to $z = d$. Thus

$$\begin{aligned} r &= [(x - \xi)^2 + (y - \eta)^2 + z^2]^{1/2} \\ &= z \left[1 + \frac{(x - \xi)^2}{z^2} + \frac{(y - \eta)^2}{z^2} \right]^{1/2} \\ &\approx z + \frac{(x - \xi)^2}{2z} + \frac{(y - \eta)^2}{2z} \end{aligned} \quad (34)$$

where we have assumed that most of the contribution to the integral comes from the domain

$$|x - \xi|, |y - \eta| \ll z$$

so that terms higher than the quadratic term can be neglected. In the denominator of Eq. (33) we may safely replace r by z so that we may write*

$$\begin{aligned} u(x, y, z) &\approx -\frac{i}{\lambda z} e^{ikz} \iint A(\xi, \eta) \\ &\times \exp \left\{ \frac{ik}{2z} [(x - \xi)^2 + (y - \eta)^2] \right\} d\xi d\eta \end{aligned} \quad (35)$$

We will apply the above formula to study the diffraction of a plane wave and also of Gaussian beam.

*For example, for $\lambda = 6 \times 10^{-5}$ cm, the factor $\cos kr$ becomes

$$\cos \left(\frac{\pi}{3} 10^5 r \right)$$

As the value of r is changed from say 60 cm to 60.00002 cm, the cosine factor will change from +1 to -0.5. This shows the rapidity with which the exponential factor will vary in the domain of integration, although the change in r is extremely small.

17.5.1 Uniform Amplitude and Phase Distribution

For such a case, at $z = 0$

$$A(\xi, \eta) = A \text{ for all values of } \xi \text{ and } \eta$$

Thus

$$u(x, y, z) = -\frac{iA}{\lambda z} e^{ikz} \int_{-\infty}^{+\infty} e^{\frac{ik}{2z} X^2} dX \int_{-\infty}^{+\infty} e^{\frac{ik}{2z} Y^2} dY$$

where $X = x - \xi$ and $Y = y - \eta$. If we now use the integral (see Appendix A)

$$\int_{-\infty}^{+\infty} e^{-\alpha x^2 + \beta x} dx = \sqrt{\frac{\pi}{\alpha}} \exp \left[\frac{\beta^2}{4\alpha} \right] \quad (36)$$

we would get

$$u(x, y, z) = -\frac{iA}{\lambda z} e^{ikz} \left[\sqrt{\frac{\pi 2z}{-ik}} \right] \left[\sqrt{\frac{\pi 2z}{-ik}} \right]$$

or,

$$u(x, y, z) \approx A e^{ikz} \quad (37)$$

as it indeed should for a uniform plane wave. This shows that the value of the proportionality constant C chosen in Eq. (22) was correct!

17.5.2 Diffraction of a Gaussian Beam

We next consider a Gaussian beam propagating along the z -direction whose amplitude distribution on the plane $z = 0$ is given by

$$A(\xi, \eta) = a \exp \left[-\frac{\xi^2 + \eta^2}{w_0^2} \right] \quad (38)$$

implying that the phase front is plane at $z = 0$. From Eq. (38) it follows that at a distance w_0 from the z -axis, the amplitude falls by a factor $1/e$ (i.e., the intensity reduces by a factor $1/e^2$). This quantity w_0 is called the *spot size* of the beam. Substituting Eq. (38) in Eq. (35) and carrying out the integration [using Eq. (36)], we obtain [see Appendix B]

$$u(x, y, z) \approx \frac{a}{(1 + i\gamma)} \exp \left[-\frac{x^2 + y^2}{w^2(z)} \right] e^{i\phi} \quad (39)$$

where

$$\gamma = \frac{\lambda z}{\pi w_0^2} \quad (40)$$

$$w(z) = w_0 [1 + \gamma^2]^{1/2} = w_0 \left[1 + \frac{\lambda^2 z^2}{\pi^2 w_0^4} \right]^{1/2} \quad (41)$$

$$\Phi = kz + \frac{k}{2R(z)} (x^2 + y^2) \quad (42)$$

$$R(z) \equiv z \left(1 + \frac{1}{\gamma^2} \right) = z \left[1 + \frac{\pi^2 w_0^4}{\lambda^2 z^2} \right] \quad (43)$$

Thus the intensity distribution is given by

$$I(x, y, z) = \frac{I_0}{1 + \gamma^2} \exp \left[-\frac{2(x^2 + y^2)}{w^2(z)} \right] \quad (44)$$

which show that the transverse intensity distribution remains Gaussian with the beam-width increasing with z which essentially implies diffraction divergence. As can be seen from Eq. (41), for small values of z , the width increases quadratically with z but for large values of $z \gg w_0^2/\lambda$, we obtain

$$w(z) = w_0 \frac{\lambda z}{\pi w_0^2} = \frac{\lambda z}{\pi w_0} \quad (45)$$

which shows that the width increases linearly with z . We define the diffraction angle as

$$\tan \theta = \frac{w(z)}{z} \approx \frac{\lambda}{\pi w_0}$$

showing that the rate of increase in the width is proportional to the wavelength and inversely proportional to the initial width of the beam. In order to get some numerical values we assume $\lambda = 0.5 \mu\text{m}$. Then, for $w_0 = 1\text{mm}$

$$2\theta \approx 0.018^\circ \quad \text{and} \quad w \approx 1.88 \text{ mm at } z = 10 \text{ m}$$

Similarly, for $w_0 = 0.25 \text{ mm}$,

$$2\theta \approx 0.073^\circ \quad \text{and} \quad w \approx 6.35 \text{ mm at } z = 2 \text{ m}$$

(see Fig. 17.9). Notice that θ increases with decrease in w_0 (smaller the size of the aperture, greater is the diffraction). Further, for a given value of w_0 , the diffraction effects decrease with λ . In Fig. 17.10 we have shown the decrease in diffraction divergence for $w_0 = 0.25 \text{ mm}$ as the wavelength is decreased from 5000 \AA to 500 \AA ; indeed as $\lambda \rightarrow 0$, $\theta \rightarrow 0$ and there is no diffraction which is the geometric optics limit. From Eq. (44) one can readily show that

$$\int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} I(x, y, z) dx dy = \frac{\pi w_0^2}{2} I_0$$

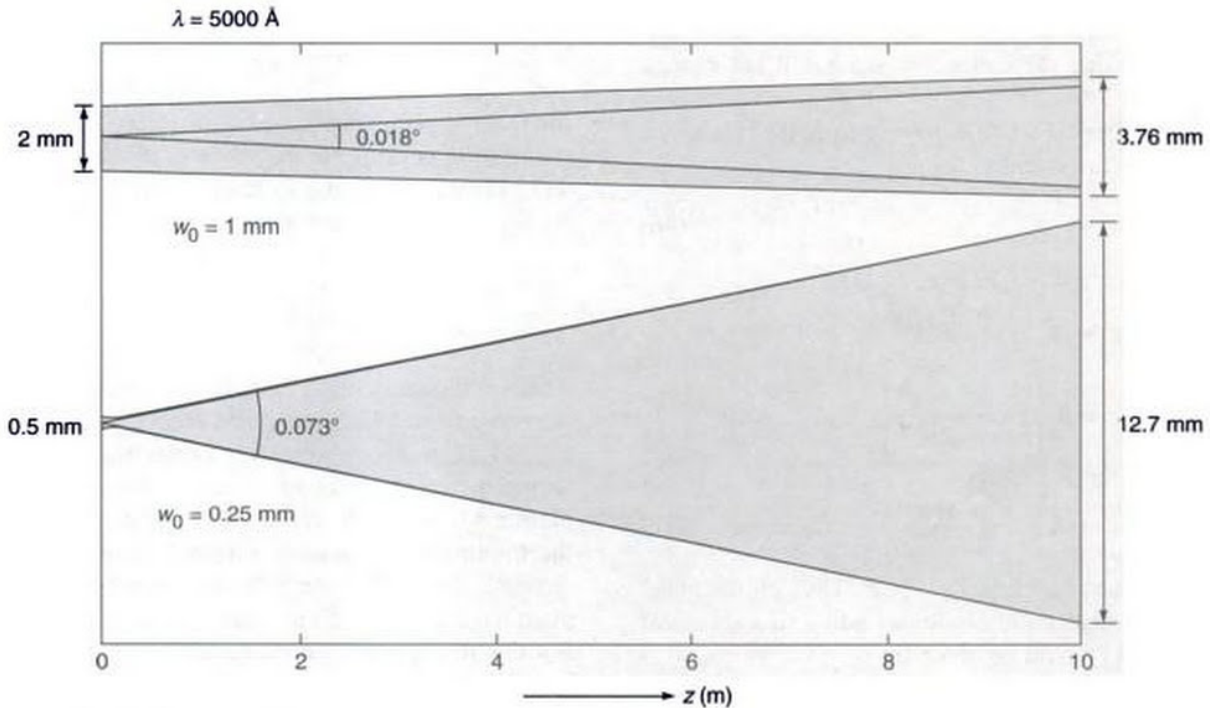


Fig. 17.9 Diffraction divergence of a Gaussian beam whose phase front is plane at $z = 0$. The figure shows the increase in the diffraction divergence as the initial spot size is decreased from 1 mm to 0.25 mm ; the wavelength is assumed to be 5000 \AA .

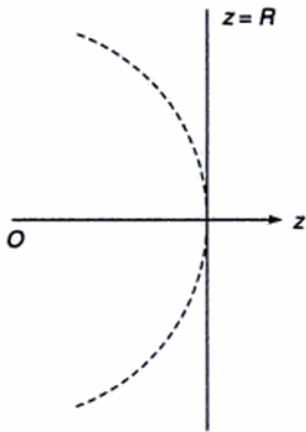


Fig. 17.11 A spherical wave diverging from the point O. The dashed curve represents a section of the spherical wavefront at a distance R from the source.

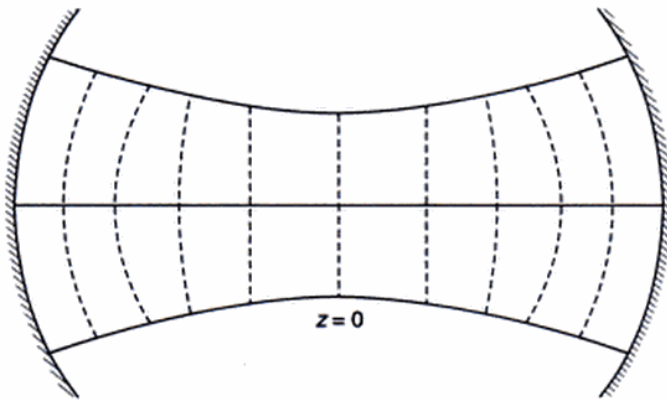


Fig. 17.12 Diffraction divergence of a Gaussian beam whose phase front is plane at $z = 0$. The dashed curves represent the phase fronts.

It should be mentioned that although in the derivation of Eq. (39) we have assumed z to be large [see Eq.(34)], Eq. (39) does give the correct field distribution even at $z = 0$.

17.6 DIFFRACTION BY A STRAIGHT EDGE

Let us consider a straight edge MN placed perpendicular to the plane of the paper and parallel to a long narrow slit S (see Fig. 17.13). We wish to calculate the intensity variation on the screen LL' . From the geometry of the arrangement it is obvious that on the screen there will be no intensity variation along the direction parallel to the length of the edge. Thus, the fringes (wherever they occur) will be straight lines parallel to the edge. We will first give a very

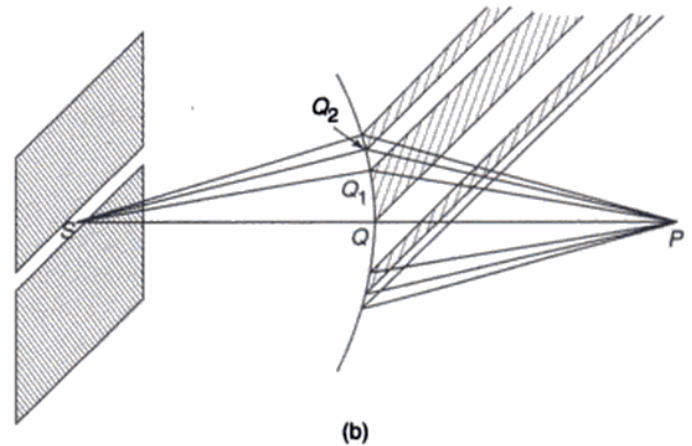
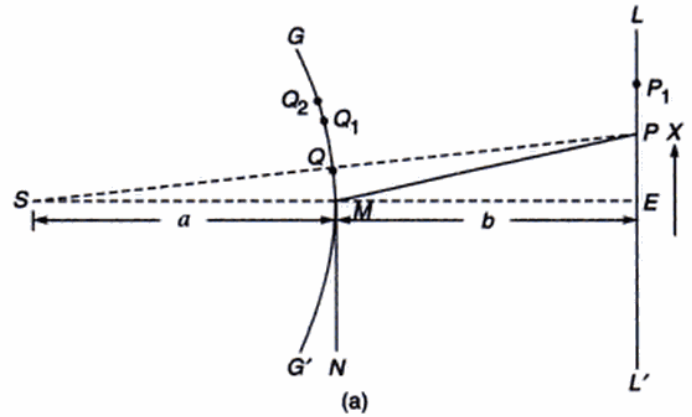


Fig. 17.13 (a) Diffraction at a straight edge. (b) Half-period strips of a cylindrical wavefront.

approximate theory based on Fresnel half-period zones; this will be followed by a more rigorous analysis.

17.6.1 Analysis Using Half Period Zones

In this section we will give a very approximate theory based on Fresnel half-period zones. The wavefront emanating from the slit is cylindrical and in order to find the amplitude at an arbitrary point P (on the screen), we draw half-period strips in the following manner: Let $GQMG'$ represent a section of the wavefront, the point Q lies on the line joining S and P . The points Q_1 and Q_2 on the wavefront are such that

$$\left. \begin{aligned} SQ_1 + Q_1P - SQP &= \frac{\lambda}{2} \\ SQ_2 + Q_2P - SQP &= \frac{2\lambda}{2}, \text{ etc.} \end{aligned} \right\} \quad (51)$$

The half-period strips will be on the surface of the cylindrical wavefront as shown in Fig. 17.13(b). However, unlike the Fresnel half-period zones, the areas of the half-period

strips will not be equal and thus the analysis becomes quite difficult. Even then one can draw the following conclusions:

- (i) Corresponding to the edge of the geometrical shadow (which is shown as E in Fig. 17.13(a)), half of the wavefront is obstructed by the edge, hence the amplitude will be given by

$$u(E) = \frac{1}{2} u_0 \quad (52)$$

where u_0 represents the amplitude that would be produced by the unobstructed wavefront (i.e., in the absence of the edge). Thus the intensity will be given by

$$I(E) = \frac{1}{4} I_0 \quad (53)$$

- (ii) Let us next assume that the point P satisfies the following relation:

$$SM + MP - SQP = \frac{\lambda}{2} \quad (54)$$

Thus only the first half-period strip of the lower part of the wavefront contributes and the resultant amplitude would approximately be

$$\frac{u_1}{2} + \frac{u_1}{4} = \frac{3u_1}{4} = \frac{3}{2} \left(\frac{u_1}{2} \right) = \frac{3}{2} u_0 \quad (55)$$

where $\frac{u_1}{2}$ is the amplitude produced by the first half-period strip in the lower portion and $\frac{u_1}{4}$ is the resultant amplitude produced by the upper half of the wavefront [see Eq.(12)]. The intensity would be $\frac{9}{4} I_0$. For a point P_1 such that

$$SM + MP_1 - SP_1 = \lambda \quad (56)$$

we will have a minimum and the resultant amplitude will be

$$\left(\frac{u_1}{2} - \frac{u_2}{2} \right) + \frac{u_1}{4} \quad (57)$$

In general, an arbitrary point P will correspond to maximum intensity if

$$SM + MP - SP = (2n + 1) \frac{\lambda}{2}; \quad n = 0, 1, 2, \dots \quad (58)$$

and a minimum if

$$SM + MP - SP = 2n \frac{\lambda}{2}; \quad n = 1, 2, \dots \quad (59)$$

Now,

$$MP = [b^2 + x^2]^{1/2} \approx b \left(1 + \frac{1}{2} \frac{x^2}{b^2} \right) = b + \frac{x^2}{2b}$$

$$SP = [(a + b)^2 + x^2]^{1/2} \approx (a + b) + \frac{x^2}{2(a + b)}$$

Hence,

$$\begin{aligned} SM + MP - SP &\approx a + b + \frac{x^2}{2b} - (a + b) - \frac{x^2}{2(a + b)} \\ &\approx \frac{a}{2(a + b)b} x^2 \end{aligned}$$

Thus, when

$$x \equiv \left[(2n + 1) \frac{b(a + b)}{a} \lambda \right]^{1/2}; \quad n = 1, 2, \dots \quad (60)$$

we will have a maximum. For example, for $a = b = 25$ cm and $\lambda = 5 \times 10^{-5}$ cm,

$$\left[\frac{b(a + b)}{a} \lambda \right]^{1/2} = 5 \times 10^{-2} \text{ cm}$$

Thus the first maximum will occur at a distance of 0.05 cm from the edge of the shadow and the second and third maxima will occur at distances of 0.0866 cm and 0.112 cm respectively. The distance between two consecutive maxima will decrease as we go away from the edge of the geometrical shadow.

Similarly, the positions of the minima are given by

$$x \approx \left[2n \frac{b(a + b)}{a} \lambda \right]^{1/2}; \quad n = 1, 2, 3, \dots \quad (61)$$

and for the above parameters they will occur at distance of 0.07 cm, 0.10 cm, etc. By determining the positions of these maxima and minima one can calculate the wavelength. The precise variation of the intensity is difficult to calculate from this analysis; a more rigorous theory will be given now.

17.6.2 More Rigorous Analysis of the Straight Edge Diffraction Pattern

Before we discuss the straight edge diffraction pattern, we introduce the Fresnel integrals.

Fresnel Integrals

The Fresnel integrals are defined by the following equations:

$$C(\tau) = \int_0^\tau \cos\left(\frac{1}{2} \pi u^2\right) du \quad (62)$$

and

$$S(\tau) = \int_0^\tau \sin\left(\frac{1}{2} \pi u^2\right) du \quad (63)$$

Since the integrands are even functions of τ , the Fresnel integrals $C(\tau)$ and $S(\tau)$ are odd functions of τ :

$$C(-\tau) = -C(\tau) \quad \text{and} \quad S(-\tau) = -S(\tau) \quad (64)$$

Further, since

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = \sqrt{\frac{\pi}{\alpha}} \quad (65)$$

we have

$$\int_{-\infty}^{\infty} e^{i\pi u^2/2} du = \sqrt{\frac{\pi}{-i\pi/2}} = \sqrt{2} e^{i\pi/4} = (1+i) \quad (66)$$

Now,

$$\begin{aligned} & \int_{-\infty}^{+\infty} \exp\left[i\frac{\pi u^2}{2}\right] du \\ &= 2 \left[\int_0^{\infty} \cos\left(\frac{1}{2}\pi u^2\right) du + i \int_0^{\infty} \sin\left(\frac{1}{2}\pi u^2\right) du \right] \\ &= 2 [C(\infty) + i S(\infty)] \end{aligned}$$

Thus, using Eq.(66), we get $C(\infty) = \frac{1}{2} = S(\infty)$.

To summarize, the Fresnel integrals have the following important properties:

$$C(\infty) = S(\infty) = \frac{1}{2}; \quad C(0) = S(0) = 0 \quad (67)$$

$$C(-\tau) = -C(\tau) \quad \text{and} \quad S(-\tau) = -S(\tau) \quad (68)$$

The values of the Fresnel integrals for typical values of τ are tabulated in Table 17.1.

Table 17.1 Table of Fresnel Integrals*

$$C(\tau) = \int_0^{\tau} \cos\left(\frac{\pi}{2} v^2\right) dv; \quad S(\tau) = \int_0^{\tau} \sin\left(\frac{\pi}{2} v^2\right) dv$$

τ	$C(\tau)$	$S(\tau)$	τ	$C(\tau)$	$S(\tau)$
0.0	0.00000	0.00000	2.6	0.38894	0.54999
0.2	0.19992	0.00419	2.8	0.46749	0.39153
0.4	0.39748	0.03336	3.0	0.60572	0.49631
0.6	0.58110	0.11054	3.2	0.46632	0.59335
0.8	0.72284	0.24934	3.4	0.43849	0.42965
1.0	0.77989	0.43826	3.6	0.58795	0.49231
1.2	0.71544	0.62340	3.8	0.44809	0.56562
1.4	0.54310	0.71353	4.0	0.49843	0.42052
1.6	0.36546	0.63889	4.2	0.54172	0.56320
1.8	0.33363	0.45094	4.4	0.43833	0.46227
2.0	0.48825	0.34342	4.6	0.56724	0.51619
2.2	0.63629	0.45570	4.8	0.43380	0.49675
2.4	0.55496	0.61969	5.0	0.56363	0.49919
			∞	0.5	0.5

*Table adapted from Ref. 5; a more detailed table (with greater accuracy) has been given there.

Figure 17.14 gives a parametric representation of the Fresnel integrals and is known as the Cornu's spiral. The horizontal and the vertical axes represent $C(\tau)$ and $S(\tau)$ respectively and the numbers written on the spiral are the values of τ . For example, as can be seen from the figure, for $\tau = 1.0$, $C(\tau) \approx 0.77989$ and $S(\tau) \approx 0.43826$.

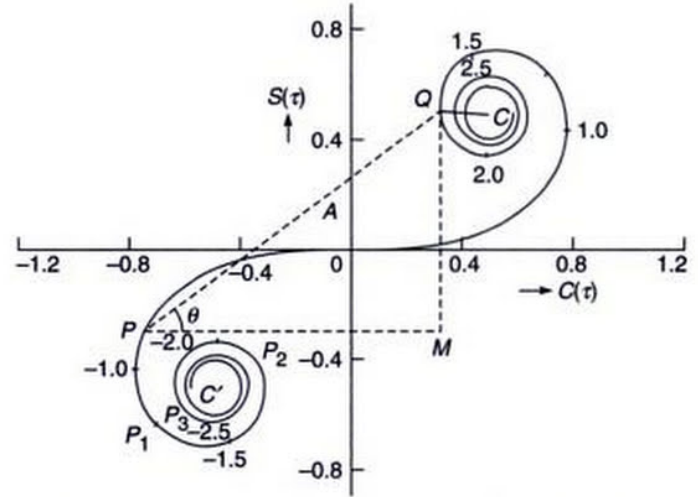


Fig. 17.14 The Cornu's spiral which is a parametric plot of $C(\tau)$ and $S(\tau)$.

We now return to the calculation of the straight edge diffraction pattern which we had qualitatively discussed in Sec.17.6.1. In this section we will make a more rigorous analysis of the diffraction of a plane wave incident normally on a straight edge (see Fig. 17.15). Once again, there will be no variation of intensity along the x -axis and, therefore, without any loss of generality, we may assume the coordinates of an arbitrary point P (on the screen) to be $(0, y)$, where the origin has been assumed to be on the edge of the geometrical shadow. If the x and y coordinates of an arbitrary point M on the plane of the straight edge are denoted by ξ and η , then

$$\begin{aligned} r &= MP = [\xi^2 + (\eta - y)^2 + d^2]^{1/2} \\ &= d \left[1 + \frac{\xi^2 + (\eta - y)^2}{d^2} \right]^{1/2} \\ &\approx d + \frac{\xi^2 + (\eta - y)^2}{2d} \end{aligned} \quad (69)$$

where d is the distance between the straight edge and the screen. On substituting the expression for r from Eq.(69) in Eq.(23), we obtain

$$u(P) \approx -\frac{i}{\lambda} \frac{A}{d} \int_{-\infty}^{\infty} d\xi \int_0^{\infty} d\eta \exp \left[ik \left\{ d + \frac{\xi^2 + (\eta - y)^2}{2d} \right\} \right] \quad (70)$$

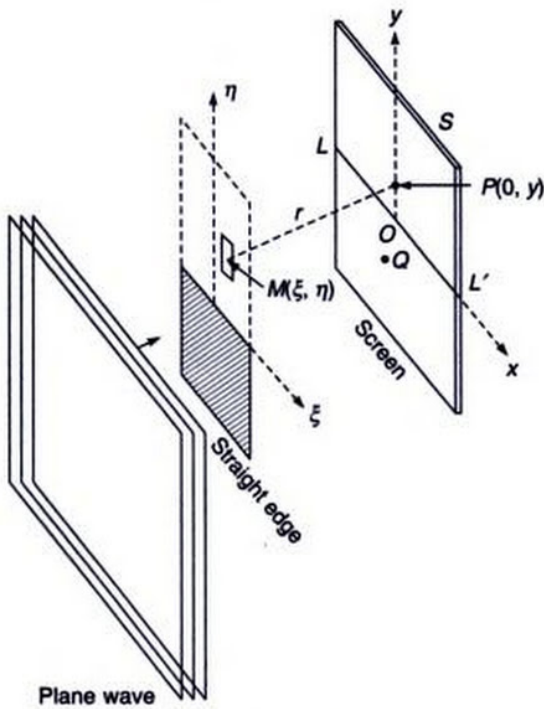


Fig. 17.15 Diffraction of a plane wave incident normally on a straight edge.

where, in the denominator of the integrand, we have replaced r by its minimum value*, d . In order to express the above expression in terms of the Fresnel integrals, we introduce two dimensionless variables u and v such that

$$\left. \begin{aligned} \frac{1}{2} \pi u^2 &= \frac{k}{2d} \xi^2 = \frac{\pi}{\lambda d} \xi^2 \\ \frac{1}{2} \pi v^2 &= \frac{k}{2d} (\eta - y)^2 = \frac{\pi}{\lambda d} (\eta - y)^2 \end{aligned} \right\} \quad \text{and}$$

Thus we may assume u and v to be defined by the following equations:

$$\left. \begin{aligned} u &= \sqrt{\frac{2}{\lambda d}} \xi \\ v &= \sqrt{\frac{2}{\lambda d}} (\eta - y) \end{aligned} \right\} \quad \text{and} \quad (71)$$

With these substitutions, Eq.(70) becomes

$$u(P) = -\frac{i}{2} u_0 \int_{-\infty}^{+\infty} \exp\left(\frac{i\pi u^2}{2}\right) du \int_{v_0}^{\infty} \exp\left(\frac{i\pi v^2}{2}\right) dv \quad (72)$$

where

$$v_0 = -\sqrt{\frac{2}{\lambda d}} y \quad (73)$$

*This is justified because in carrying out the integration, only a small region around the point $r = d$ contributes; the contribution due to far-off points is small because of the rapid oscillations of the exponential term in the integrand (see also footnote in Sec. 17.5).

and

$$u_0 = A e^{ikd}$$

represents the field at the point P in the absence of the straight edge. In order to calculate the intensity distribution we use the Fresnel integrals; thus

$$\begin{aligned} \int_{-\infty}^{+\infty} \exp\left[i\frac{\pi u^2}{2}\right] du &= 2 [C(\infty) + S(\infty)] \\ &= (1 + i) \end{aligned} \quad (74)$$

Further,

$$\begin{aligned} \int_{v_0}^{\infty} \exp\left(\frac{i\pi v^2}{2}\right) dv &= \int_0^{\infty} \cos\left(\frac{\pi}{2} v^2\right) dv - \int_0^{v_0} \cos\left(\frac{\pi}{2} v^2\right) dv \\ &\quad + i \left[\int_0^{\infty} \sin\left(\frac{\pi}{2} v^2\right) dv - \int_0^{v_0} \sin\left(\frac{\pi}{2} v^2\right) dv \right] \\ &= \left[\frac{1}{2} - C(v_0) \right] + i \left[\frac{1}{2} - S(v_0) \right] \end{aligned} \quad (75)$$

Substituting in Eq.(72), we obtain

$$\begin{aligned} u(P) &= -\frac{i}{2} u_0 (1 + i) \left[\left\{ \frac{1}{2} - C(v_0) \right\} + i \left\{ \frac{1}{2} - S(v_0) \right\} \right] \\ &= \frac{1-i}{2} u_0 \left[\left\{ \frac{1}{2} - C(v_0) \right\} + i \left\{ \frac{1}{2} - S(v_0) \right\} \right] \end{aligned} \quad (76)$$

It is of interest to note that a large value of y corresponds to a point which is very far above the edge of the geometrical shadow. For such a point v_0 would tend to $-\infty$ [see Eq.(73)] and we would obtain

$$\begin{aligned} u(P) &= \frac{1-i}{2} u_0 \left[\left(\frac{1}{2} + \frac{1}{2} \right) + i \left(\frac{1}{2} + \frac{1}{2} \right) \right] \\ &= u_0 \end{aligned} \quad (77)$$

Thus, as expected, the amplitude at such a point is the same as that in the absence of the edge. This also justifies the value of the constant given by Eq.(22). On the other hand, when the point P is deep inside the geometrical shadow (i.e., when $y \rightarrow -\infty$ and hence $v_0 \rightarrow \infty$), we obtain

$$C(v_0) = S(v_0) \rightarrow \frac{1}{2}$$

giving

$$u(P) \rightarrow 0$$

as it should indeed be. The intensity distribution corresponding to Eq.(76) would be given by

$$I(P) = \frac{1}{2} I_0 \left[\left\{ \frac{1}{2} - C(v_0) \right\}^2 + \left\{ \frac{1}{2} - S(v_0) \right\}^2 \right] \quad (78)$$

If the point P is such that it lies on the edge of the geometrical shadow (i.e., on the line LL' (see Fig. 17.15) then $y = 0$ and hence $v_0 = 0$; thus

$$I(P) = \frac{1}{2} I_0 \left[\frac{1}{4} + \frac{1}{4} \right] = \frac{1}{4} I_0 \quad (79)$$

where we have used the fact that $C(0) = S(0) = 0$. Thus the intensity on the edge of the geometrical shadow is $1/4^{\text{th}}$ of the intensity that would have been in the absence of the edge [see also Eq.(53)]. In order to determine the field at an arbitrary point P , we may use Table 17.1 to calculate the RHS of Eq.(78). The intensity variation is plotted in Fig. 17.16 from which one can make the following observations:

- (i) Figure 17.16 represents a universal curve, i.e., for given values of λ and d , one simply has to calculate v_0 as the observation point moves along the y -axis. For example, the first three maxima occur at

$$\left. \begin{array}{l} v_0 = -1.22 \text{ with } I \approx 1.37 I_0 \\ v_0 = -2.34 \text{ with } I \approx 1.20 I_0 \\ v_0 = -3.08 \text{ with } I \approx 1.15 I_0 \end{array} \right\} \text{maxima}$$

Similarly, the first three minima occur at

$$\left. \begin{array}{l} v_0 = -1.87 \text{ with } I \approx 0.778 I_0 \\ v_0 = -2.74 \text{ with } I \approx 0.843 I_0 \\ v_0 = -3.39 \text{ with } I \approx 0.872 I_0 \end{array} \right\} \text{minima}$$

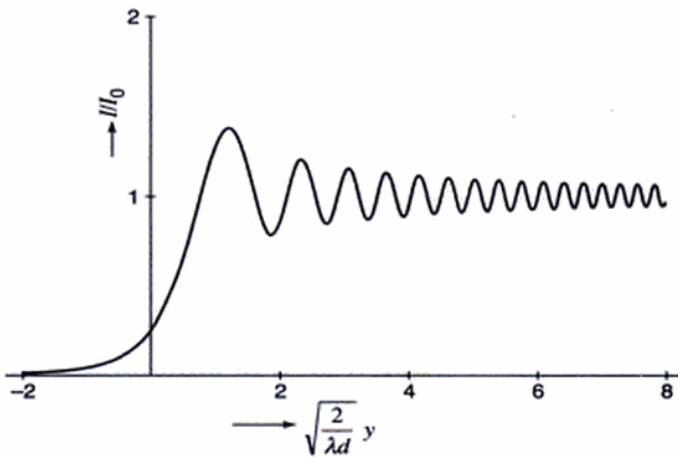


Fig. 17.16 The intensity variation corresponding to the straight edge diffraction pattern.

Thus, as we go inside the geometrical shadow, the intensity modulation decreases (see Fig. 17.16).

- (ii) For a given experimental set-up, the determination of the positions of maxima and minima is quite straightforward. For example, for $\lambda = 6 \times 10^{-5}$ cm and $d = 120$ cm

$$y = -\sqrt{\frac{\lambda d}{2}} v_0 = -0.06 v_0 \text{ cm}$$

Thus the first three maxima will occur at

$$y \approx 0.732, 1.404 \text{ and } 1.848 \text{ mm}$$

respectively. Similarly, the first three minima will occur at

$$y \approx 1.122, 1.644 \text{ and } 2.034 \text{ mm}$$

respectively. These results may be compared with those obtained in Sec. 17.6.1.

- (iii) As we go inside the geometrical shadow the intensity monotonically decreases to zero.
(iv) One could have also studied the intensity variation directly from the Cornu's spiral (see Fig. 17.14). This is due to the fact that associated with the Cornu's spiral, we have the following interesting property: let us write

$$[C(\tau_2) - C(\tau_1)] + i[S(\tau_2) - S(\tau_1)] \equiv A e^{i\theta} \quad (80)$$

Thus

$$[C(\tau_2) - C(\tau_1)] = A \cos \theta$$

and

$$[S(\tau_2) - S(\tau_1)] = A \sin \theta$$

Let the points P and Q on the Cornu's spiral (see Fig. 17.14) correspond to $\tau = \tau_1$ and $\tau = \tau_2$ respectively. It is obvious that

$$PM = [C(\tau_2) - C(\tau_1)] = A \cos \theta$$

and

$$QM = [S(\tau_2) - S(\tau_1)] = A \sin \theta$$

Thus the length of the line joining the points P and Q will be A and the angle that the line makes with the abscissa will be θ . In order to use the Cornu's spiral we rewrite Eq. (76):

$$u = \frac{1-i}{2} u_0 \left[\left\{ \frac{1}{2} - C(v_0) \right\} + i \left\{ \frac{1}{2} - S(v_0) \right\} \right]$$

Let us first consider a point of observation Q in the geometrical shadow region. Consequently v_0 will be positive. Let the point Q on the spiral (see Fig. 17.14)

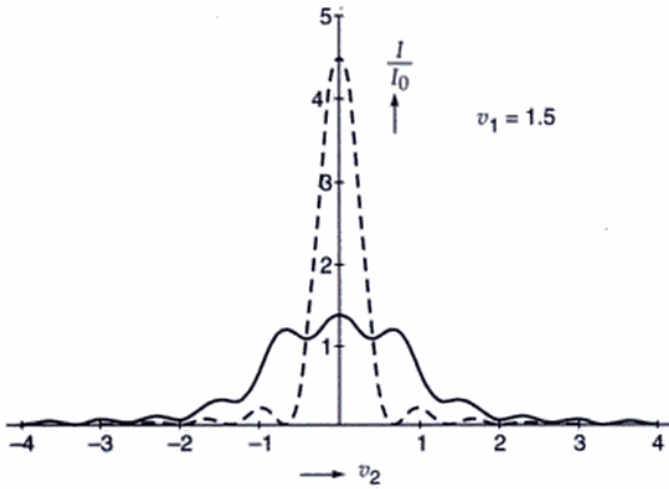


Fig. 17.21 The intensity distribution produced by diffraction of a plane wave by a long narrow slit corresponding to $v_1 = 1.5$. The dashed curves correspond to Eq. (86).

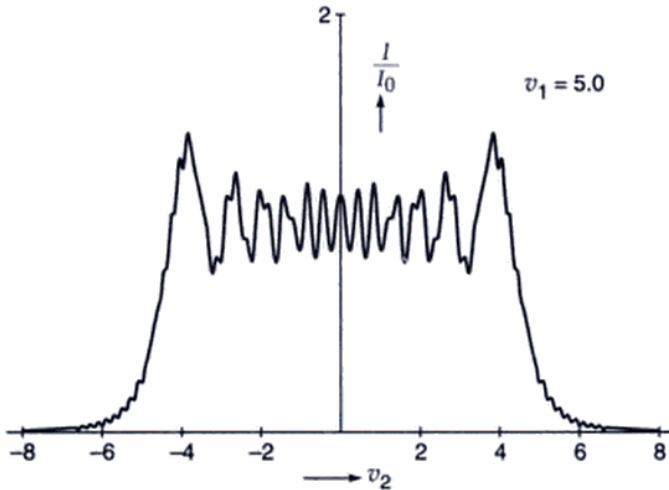


Fig. 17.22 The intensity distribution produced by diffraction of a plane wave by a long narrow slit corresponding to $v_1 = 5.0$.

the diffraction pattern is essentially of the Fraunhofer type. In order to show this explicitly we notice that

$$v_2 = \sqrt{\frac{2}{\lambda d}} y = \sqrt{\frac{2d}{\lambda}} \frac{y}{d} \approx \sqrt{\frac{2d}{\lambda}} \theta \quad (85)$$

where θ represents the angle of diffraction (see Fig. 17.23). Clearly, in the Fraunhofer region since d is very large, the value of v_2 will also be very large and thus we must look for expressions of the Fresnel integrals in the limit of $v \rightarrow \infty$. Now, we may write

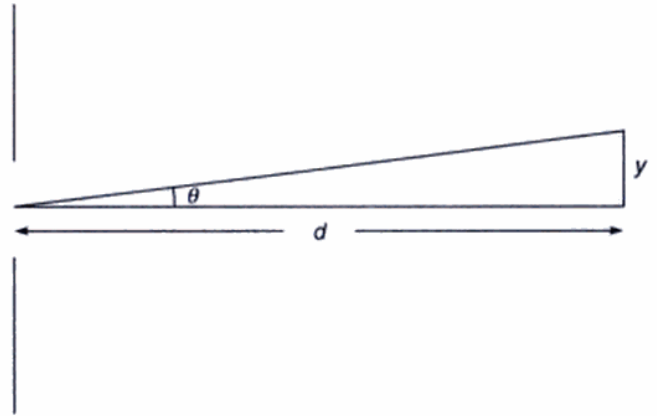


Fig. 17.23 In the Fraunhofer region, d is very large.

$$\begin{aligned} C(v) &= \int_0^v \cos \frac{\pi}{2} v^2 dv \\ &= \int_0^\infty \cos \frac{\pi}{2} v^2 dv - \int_v^\infty \cos \frac{\pi}{2} v^2 dv \\ &= \frac{1}{2} - \int_v^\infty \frac{1}{\pi v} \cos \left(\frac{\pi}{2} v^2 \right) \pi v dv \\ &= \frac{1}{2} - \frac{1}{\pi v} \sin \left(\frac{\pi}{2} v^2 \right) \Big|_v^\infty + \int_v^\infty \frac{1}{\pi v^2} \sin \left(\frac{\pi}{2} v^2 \right) dv \\ &\approx \frac{1}{2} + \frac{1}{\pi v} \sin \left(\frac{\pi}{2} v^2 \right) \end{aligned}$$

where we have neglected terms which would be of order $1/v^3$. Similarly

$$S(v) = \frac{1}{2} - \frac{1}{\pi v} \cos \left(\frac{\pi}{2} v^2 \right)$$

Since v_2 is large and v_1 is small, we have

$$\begin{aligned} C(v_2 + v_1) - C(v_2 - v_1) &= \left[\frac{1}{2} + \frac{1}{\pi v_2} \sin \frac{\pi}{2} (v_2 + v_1)^2 \right] \\ &\quad - \left[\frac{1}{2} + \frac{1}{\pi v_2} \sin \frac{\pi}{2} (v_2 - v_1)^2 \right] \\ &\approx \frac{2}{\pi v_2} \cos \frac{\pi}{2} (v_2^2 + v_1^2) \sin \pi v_1 v_2 \end{aligned}$$

Similarly,

$$S(v_2 + v_1) - S(v_2 - v_1) \approx \frac{2}{\pi v_2} \sin \frac{\pi}{2} (v_2^2 + v_1^2) \sin (\pi v_1 v_2)$$

Thus, in the Fraunhofer limit, Eq.(84) becomes

$$I(P) = \frac{1}{2} I_0 \left[\frac{4}{\pi^2 v_2^2} \sin^2(\pi v_1 v_2) \right] \\ = I_{00} \frac{\sin^2 \beta}{\beta^2} \quad (86)$$

where

$$I_{00} = 2 I_0 v_1^2$$

and

$$\beta = \pi v_1 v_2 = \frac{\pi}{\lambda d} b y \approx \frac{\pi b}{\lambda} \theta \quad (87)$$

and

$$\theta \approx \frac{y}{d} \quad (88)$$

represents the diffraction angle. Equation (86) shows that the intensity distribution is indeed of the Fraunhofer type (see Sec. 16.2). In Figs 17.19 – 17.22 the dotted curves correspond to Eq.(86) and one can see that the intensity distribution is almost of the Fraunhofer type for $v_1 \leq 0.5$.

17.8 FRAUNHOFER DIFFRACTION BY AN APERTURE

We now use the method developed in Sections 17.5 and 17.6 to study the Fraunhofer diffraction pattern produced by a plane wave incident normally on an aperture as shown in Fig. 17.24. The field at the point P will again be given by Eq.(23)

$$u(P) \approx -\frac{i A}{\lambda} \iint \frac{e^{ikr}}{r} d\xi d\eta \quad (89)$$

where $r = PM$, M representing an arbitrary point on the aperture. If the coordinates of the points M and P are (ξ, η) ,

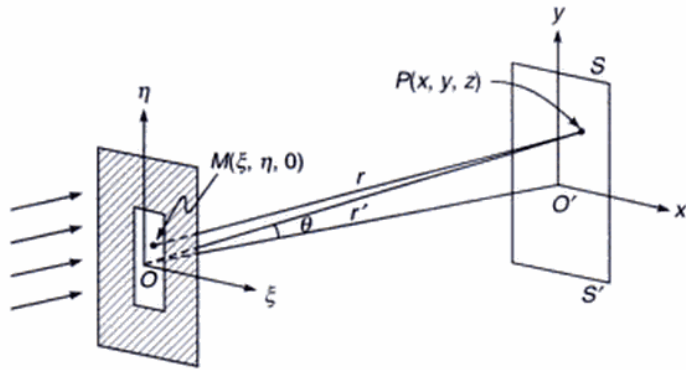


Fig. 17.24 Diffraction of a plane wave incident on an aperture (shown rectangular in the figure).

0) and (x, y, z) then

$$r = PM = [(x - \xi)^2 + (y - \eta)^2 + z^2]^{1/2} \\ = r'^2 \left[1 - \frac{2(x\xi + y\eta)}{r'^2} + \frac{\xi^2 + \eta^2}{r'^2} \right]^{1/2} \\ \approx r' - \frac{(x\xi + y\eta)}{r'} \quad (90)$$

where $r' = \sqrt{x^2 + y^2 + z^2}$ represents the distance of the point P from a conveniently chosen origin O . In writing Eq.(90) we have neglected the terms of order $1/r'^2$; this corresponds to Fraunhofer diffraction. Substituting for r from Eq.(90) in Eq.(89) we obtain

$$u(P) \approx C \iint e^{-ik(l\xi + m\eta)} d\xi d\eta \quad (91)$$

where $C = -\frac{i A e^{ikr'}}{\lambda r'}$

and $l = x/r'$ and $m = y/r'$ represent the direction cosines of the diffracted ray.

17.8.1 Rectangular Aperture

For a rectangular aperture (of dimension $a \times b$) we obtain

$$u(P) = C \int_{-b/2}^{+b/2} \int_{-a/2}^{+a/2} e^{-ik(l\xi + m\eta)} d\xi d\eta \quad (92)$$

where we have chosen the origin to be at the centre of the rectangular aperture. Since

$$\int_{-b/2}^{+b/2} e^{-ikl\xi} d\xi = \frac{1}{-ikl} e^{-ikl\xi} \Big|_{-b/2}^{+b/2} \\ = \frac{2}{kl} \frac{e^{ikbl/2} - e^{-ikbl/2}}{2i} \\ = \frac{2}{kl} \sin \frac{kbl}{2} = \frac{1}{\tau} \sin b\tau \\ = b \frac{\sin \beta}{\beta} \quad (93)$$

where

$$\beta = b\tau = \frac{kbl}{2} = \frac{\pi b \sin \theta}{\lambda} \quad (94)$$

$$l = \frac{x}{r'} \equiv \sin \theta$$

where $\zeta = k \rho \sin \theta$ and use has made of the following well-known relation:

$$J_0(\zeta) = \frac{1}{2\pi} \int_0^{2\pi} e^{\pm i\zeta \cos \phi} d\phi \quad (105)$$

If we further use the relation

$$\frac{d}{d\zeta} [\zeta J_1(\zeta)] = \zeta J_0(\zeta) \quad (106)$$

then Eq. (104) becomes

$$\begin{aligned} u(P) &= \frac{2\pi C}{(k \sin \theta)^2} [\zeta J_1(\zeta)] \Big|_0^{k a \sin \theta} \\ &= u_0 \left[\frac{2J_1(v)}{v} \right] \end{aligned}$$

where $v = k a \sin \theta$. Thus the intensity distribution would be given by

$$I(P) = I_0 \left[\frac{2J_1(v)}{v} \right]^2 \quad (107)$$

This is the famous Airy pattern which has been discussed in Sec. 16.3.

We have already mentioned that the diffraction pattern (in the plane SS') will consist of concentric rings with their centers at the point O' (see Fig. 17.27). If $F(r)$ represents the fractional energy contained in a circle of radius r , then

$$F(r) = \frac{\int_0^r I(\sigma) 2\pi \sigma d\sigma}{\int_0^\infty I(\sigma) 2\pi \sigma d\sigma} \quad (108)$$

where $I(\sigma) 2\pi \sigma d\sigma$ would be proportional to the energy contained in the annular region whose radii lie between σ and $\sigma + d\sigma$. Clearly

$$\frac{\sigma}{r'} = \sin \theta \quad (109)$$

Since $v = k a \sin \theta$, we obtain

$$\sigma = \frac{r'}{k a} v \quad (110)$$

Thus Eq. (108) becomes

$$F(r) = \frac{\int_0^v \left[\frac{2J_1(v)}{v} \right]^2 v dv}{\int_0^\infty \left[\frac{2J_1(v)}{v} \right]^2 v dv} \quad (111)$$

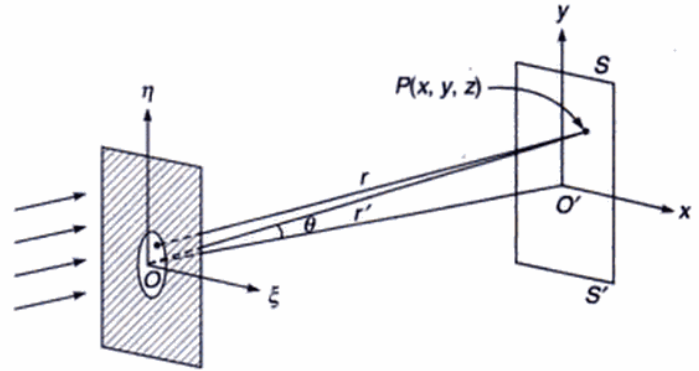


Fig. 17.27 Diffraction of a plane wave incident on a circular aperture.

where we have used Eq. (107) for the intensity distribution. Now*,

$$\begin{aligned} \frac{J_1^2(v)}{v} &= J_1(v) \left[J_0(v) - \frac{dJ_1(v)}{dv} \right] \\ &= - \left[J_0(v) \frac{dJ_0(v)}{dv} + J_1(v) \frac{dJ_1(v)}{dv} \right] \\ &= - \frac{1}{2} \frac{d}{dv} [J_0^2(v) + J_1^2(v)] \end{aligned} \quad (112)$$

Thus

$$F(r) = \frac{J_0^2(v) + J_1^2(v) \Big|_0^v}{J_0^2(v) + J_1^2(v) \Big|_0^\infty} = 1 - J_0^2(v) - J_1^2(v) \quad (113)$$

The above function is plotted in Fig. 17.28; one can deduce from the curve that about 84% of light is contained within the circle bounded by the first dark ring and about 91% of the light is contained in the circle bounded by the first two dark rings, etc.

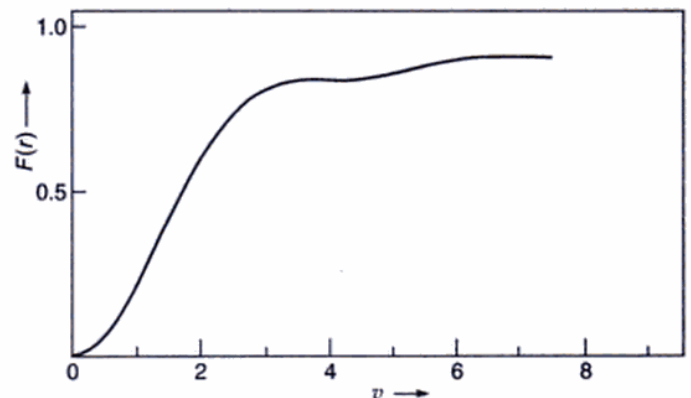


Fig. 17.28 The fractional energy contained in a circle of radius r .

*See any book on mathematical physics for the recurrence relations of Bessel functions.

SUMMARY

- The underlying principle in the theory of diffraction is the Huygens—Fresnel principle according to which: *Each point on a wavefront is a source of secondary disturbance and the secondary wavelets emanating from different points mutually interfere.*
- For a plane wave incident normally on a circular aperture of radius a , the intensity variation on an axial point P is given by

$$I = I_0 \sin^2 \frac{p\pi}{2}$$

where

$$p \approx \frac{a^2}{\lambda d}$$

λ is the wave length and d is the distance of the point P from the center of the circular aperture. The quantity p is known as the Fresnel number of the aperture. When $p = 1, 3, 5, 7, \dots$ we have maximum intensity and the circular aperture will contain (with respect to the point P) odd number of Fresnel half period zones and when $p = 2, 4, 6, 8, \dots$ we have minimum intensity and the circular aperture will contain even number of half period zones.

- If instead of the circular aperture we have opaque disc, then we would always obtain a bright spot on the axis behind the disc; this is called the 'Poisson spot'.
- For a Gaussian beam (whose phase front is plane at $z = 0$), the variation of the spot size is given by

$$w(z) \approx w_0 \left[1 + \frac{\lambda^2 z^2}{\pi^2 w_0^4} \right]^{1/2}$$

where w_0 is the spot size at $z = 0$. For large values of z :

$$w(z) \approx \frac{\lambda z}{\pi w_0}$$

which shows that the width increases linearly with z . We define the diffraction angle as

$$\tan \theta = \frac{w(z)}{z} \approx \frac{\lambda}{\pi w_0}$$

showing that the rate of increase in width is proportional to the wavelength and inversely proportional to the initial width of the beam; this is characteristic of diffraction. The corresponding radius of curvature of the wavefront is given by

$$R(z) \approx z \left[1 + \frac{\pi^2 w_0^4}{\lambda^2 z^2} \right]$$

- For a plane wave incident normally on a straight edge, the intensity variation on a screen (at a distance d from the straight edge) is given by

$$I = \frac{1}{2} I_0 \left[\left\{ \frac{1}{2} - C(v_0) \right\}^2 + \left\{ \frac{1}{2} - S(v_0) \right\}^2 \right]$$

where I_0 is the intensity in the absence of the straight edge,

$$v_0 = -\sqrt{\frac{2}{\lambda d}} y$$

y being the distance from the edge of the geometrical shadow and

$$C(x) = \int_0^x \cos\left(\frac{1}{2} \pi u^2\right) du$$

$$\text{and } S(x) = \int_0^x \sin\left(\frac{1}{2} \pi u^2\right) du$$

are known as Fresnel integrals. The intensity monotonically goes to zero as we go deep inside the geometrical shadow. As we move away from the edge of the geometrical shadow to the illuminated region, one obtains maxima at $v_0 \approx -1.22$ ($I \approx 1.37 I_0$), -2.34 ($I \approx 1.20 I_0$), -3.08 ($I \approx 1.15 I_0$) ... and minima at $v_0 \approx -1.87$ ($I \approx 0.78 I_0$), -2.74 ($I \approx 0.84 I_0$), -3.39 ($I \approx 0.87 I_0$), ...

- For a plane wave incident normally on a long narrow slit of width b , the intensity variation on a screen (at a distance d from the slit) is given by

$$I = \frac{1}{2} I_0 \left[\{C(v_2 + v_1) - C(v_2 - v_1)\}^2 + \{S(v_2 + v_1) - S(v_2 - v_1)\}^2 \right]$$

where

$$v_1 = \sqrt{\frac{2}{\lambda d}} \frac{b}{2}; v_2 = \sqrt{\frac{2}{\lambda d}} y$$

and y is the distance from the midpoint of the edges of the geometrical shadow. As v_1 becomes large, we obtain the intensity distribution corresponding to two straight edges and for $v_1 \rightarrow 0$ we get the Fraunhofer diffraction pattern.

- In the Airy pattern (discussed in the previous chapter) about 84% of light is contained within the circle bounded by the first dark ring and about 91% of the light is contained in the circle bounded by the first two dark rings, etc.

PROBLEMS

- Consider a plane wave of wavelength 6×10^{-5} cm incident normally on a circular aperture of radius 0.01 cm. Calculate the positions of the brightest and the darkest points on the axis.
- What would happen if the circular aperture in Problem 17.1 is replaced by a circular disc of the same radius?

17.3 A plane wave ($\lambda = 6 \times 10^{-5}$ cm) is incident normally on a circular aperture of radius a .

- (a) Assume $a = 1$ mm. Calculate the values of z (on the axis) for which maximum intensity will occur. Plot the intensity as a function of z and interpret physically. Repeat the calculations for $\lambda = 5 \times 10^{-5}$ cm and discuss chromatic aberration of a zone plate.
- (b) Assume $z = 50$ cm. Calculate the values of a for which minimum intensity will occur on the axial point. Plot the intensity variation as a function of a and interpret physically.

17.4 Consider a circular aperture of diameter 2 mm illuminated by a plane wave. The most intense point on the axis is at a distance of 200 cm from the aperture. Calculate the wavelength.

[Ans. 5×10^{-5} cm]

17.5 If a zone-plate has to have a principle focal length of 50 cm corresponding to $\lambda = 6 \times 10^{-5}$ cm, obtain an expression for the radii of different zones. What would be its principle focal length for $\lambda = 5 \times 10^{-5}$ cm?

[$(\sqrt{0.3n})$ mm, 60 cm]

17.6 In a zone-plate, the second, fourth, sixth... zones are blackened; what would happen if instead the 1st, 3rd, 5th, etc., zones were blackened?

17.7 (a) A plane wave is incident normally on a straight edge (see Fig. 17.29). Show that the field at an arbitrary point P is given by

$$u(P) = \frac{1-i}{2} u_0 \left[\left\{ \frac{1}{2} - C(v_0) \right\} + i \left\{ \frac{1}{2} - S(v_0) \right\} \right]$$

where $v_0 = -\sqrt{\frac{2}{\lambda d}} y$.

- (b) Assume $\lambda_0 = 5000 \text{ \AA}$ and $d = 100$ cm. Write approximately the values of I/I_0 at the points O , P ($y = 0.5$ mm), Q ($y = 1$ mm) and R ($y = -1$ mm) where O is at the edge of the geometrical shadow.

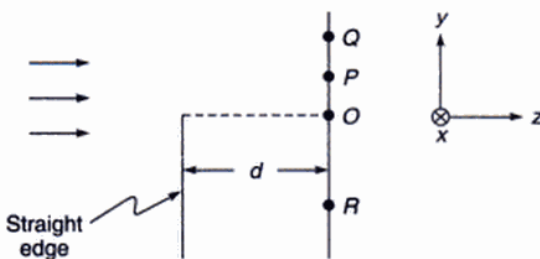


Fig. 17.29

17.8 Consider a straight edge being illuminated by a parallel beam of light with $\lambda = 6 \times 10^{-5}$ cm. Calculate the positions of the first two maxima and minima on a screen at a distance of 50 cm from the edge.

17.9 In a straight edge diffraction pattern, one observes that the most intense maximum occurs at a distance of 1 mm from the edge of the geometrical shadow. Calculate the wavelength of light, if the distance between the screen and the straight edge is 300 cm.

[Ans. $\approx 4480 \text{ \AA}$]

17.10 In a straight edge diffraction pattern, if the wavelength of the light used is 6000 \AA and if the distance between the screen and the straight edge is 100 cm, calculate the distance between the most intense maximum and the next maximum. Find approximately the distance in centimeters inside the geometrical shadow where $I/I_0 = 0.1$.

17.11 Consider a plane wave falling normally on a narrow slit of width 0.5 mm. If the wavelength of light is 6×10^{-5} cm, calculate the distance between the slit and the screen so that the value of v_1 would be 0.5, 1.0, 1.5 and 5.0 (see Fig. 17.19 – 17.22). Discuss the transition to the Fraunhofer region.

17.12 Consider the Fresnel diffraction pattern produced by a plane wave incident normally on a slit of width b . Assume $\lambda = 5 \times 10^{-5}$ cm, $d = 100$ cm. Using Table 17.1, approximately calculate the intensity values (for $b = 0.1$ cm) at $y = 0, \pm 0.05$ cm, ± 0.1 cm. Repeat the analysis for $b = 5$ cm.

17.13 In Sec. 17.9 we obtained the diffraction pattern of a circular aperture of radius a . Obtain the diffraction pattern of an annular aperture bounded by circles of radii a_1 and a_2 ($a_2 > a_1$).

[Hint: The integration limits of ρ in Eq. (103) must be a_1 and a_2]

17.14 Consider a rectangular aperture of dimensions $0.2 \text{ mm} \times 0.3 \text{ mm}$. Obtain the positions of the first few maxima and minima in the Fraunhofer diffraction pattern along directions parallel to the length and breadth of the rectangle. Assume $\lambda = 5 \times 10^{-5}$ cm and that the diffraction pattern is produced at the focal plane of a lens of focal length 20 cm.

17.15 The Fraunhofer diffraction pattern of a circular aperture (of radius 0.5 mm) is observed on the focal plane of a convex lens of focal length 20 cm. Calculate the radii of the first and the second dark rings. Assume $\lambda = 5.5 \times 10^{-5}$ cm.

[Ans. 0.13 mm, 0.25 mm]

17.16 In Problem 17.15, calculate the area of the patch (on focal plane) which will contain 95% of the total energy.

- 17.17** The output of a He-Ne laser ($\lambda = 6328 \text{ \AA}$) can be assumed to be Gaussian with plane phase front. For $w_0 = 1 \text{ mm}$ and $w_0 = 0.2 \text{ mm}$, calculate the beam diameter at $z = 20 \text{ m}$. Repeat the calculation for $\lambda = 5000 \text{ \AA}$ and interpret the results physically.
- 17.18** A Gaussian beam is coming out of a laser. Assume $\lambda = 6000 \text{ \AA}$ and that at $z = 0$, the beam width is 1 mm and the phase front is plane. After traversing 10 m through vacuum what will be (a) the beam width and (b) the radius of curvature of the phase front.
- 17.19** A plane wave of intensity I_0 is incident normally on a circular aperture as shown in Fig. 17.30. What will be the intensity on the axial point P ?

[Hint: You may use Eq. (25)]

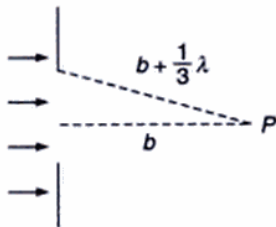


Fig. 17.30

- 17.20** Show that a phase variation of the type

$$\exp \left[ikz + \frac{ik(x^2 + y^2)}{2R(z)} \right]$$

represents a diverging spherical wave of radius R .

- 17.21** Consider a resonator consisting of a plane mirror and a concave mirror of radius of curvature R (see

Fig. 17.31). Assume $\lambda = 1 \mu\text{m}$, $R = 100 \text{ cm}$ and the distance between the 2 mirrors to be 50 cm . Calculate the spot size of the Gaussian beam.

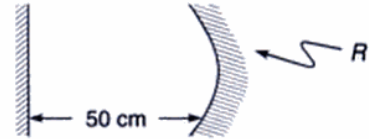


Fig. 17.31

- 17.22** The output of a semiconductor laser can be approximately described by a Gaussian function with two different widths along the transverse (w_T) and lateral (w_L) directions as

$$\psi(x, y) = A \exp \left(-\frac{x^2}{w_L^2} - \frac{y^2}{w_T^2} \right)$$

where x and y represent axes parallel and perpendicular to the junction plane. Typically $w_T \approx 0.5 \mu\text{m}$ and $w_L = 2 \mu\text{m}$. Discuss the far field of this beam (see Fig. 17.32).

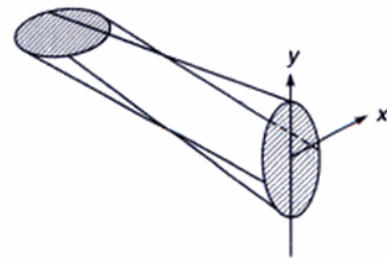


Fig. 17.32

REFERENCES AND SUGGESTED READINGS

1. R. Baierlein, *Newton to Einstein: The Trail of light*, Cambridge University Press, 1992.
2. P. M. Rinard, 'Large scale diffraction patterns from circular objects', *American Journal of Physics*, Vol. 44, 70, 1976.
3. M. Born and E. Wolf, *Principles of Optics*, Cambridge University Press, 2000.
4. A. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978.
5. M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables*, Applied Mathematics Series, Vol. 55; National Bureau of Standards, Washington (1964).

Chapter 18

Holography*

*The electron microscope was to produce the interference figure between the object beam and the coherent background, that is to say the non-diffracted part of the illuminating beam. This interference pattern I called a **hologram**, from the Greek word **holos** – the whole, because it contained the whole information. The hologram was then reconstructed with light, in an optical system which corrected the aberrations of the electron optics.*

– Dennis Gabor in his Nobel lecture** December 11, 1971

Important Milestones

- | | |
|------|--|
| 1948 | Dennis Gabor discovered the principle of holography |
| 1960 | The first successful operation of a laser device by Theodore Maiman |
| 1962 | Off-axis technique of holography by Leith and Upatnieks |
| 1962 | Denisyuk suggested the idea of three-dimensional holograms based on thick photoemulsion layers. His holograms can be reconstructed in ordinary sun light. These holograms are called Lippmann – Bragg holograms. |
| 1964 | Leith and Upatnieks pointed out that a multicolour image can be produced by a hologram recorded with three suitably chosen wavelengths. |
| 1969 | Benton invented 'Rainbow Holography' for display of holograms in white light. This was a vital step to make holography suitable for display applications. |

18.1 INTRODUCTION

A photograph represents a two-dimensional recording of a three-dimensional scene. What is recorded is the intensity distribution that prevailed at the plane of the photograph when it was exposed. The light sensitive medium is sensitive only to the intensity variations and hence while recording a photograph, the phase distribution which prevailed at the plane of the photograph is lost. Since only the intensity pattern has been recorded, the three-dimensional character (e.g., parallax) of the object scene is lost. Thus one cannot change the perspective of the image in the photograph by viewing it from a different angle or one cannot refocus any unfocussed part of the image in the photograph. Holography is a method invented by Dennis Gabor in 1947, in which one not only records the amplitude but also the phase of the light wave; this is done by using interferometric tech-

niques. Because of this, the image produced by the technique of holography has a true three-dimensional form. Thus, as with the object, one can change one's position and view a different perspective of the image or one can focus at different distances. The capability to produce images as true as the object itself is what is responsible for the wide popularity gained by holography.

The basic technique in holography is the following: In the recording of the hologram, one superimposes on the object wave another wave called the reference wave and the photographic plate is made to record the resulting interference pattern (see Fig. 18.1). The reference wave is usually a plane wave. This recorded interference pattern forms the hologram and (as will be shown) contains information not only about the amplitude but also about the phase of the object wave. Unlike a photograph, a hologram has little resemblance with the object; in fact, information

*A portion of this chapter is based on the unpublished lecture notes of Professor K. Thyagarajan.

**Dennis Gabor received the 1971 Nobel Prize in Physics for discovering the *principles of holography*; the original paper of Gabor appeared in 1948 [see Ref. 1]. Gabor's Nobel lecture entitled *Holography, 1948 – 1971* is non-mathematical and full of beautiful illustrations; it is reprinted in Ref. 2.

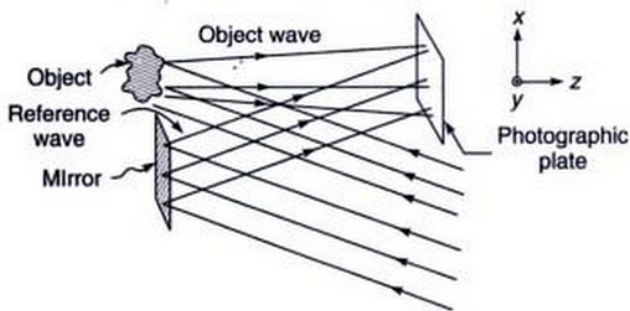


Fig. 18.1 Recording of a hologram.

about the object is coded into the hologram. To view the image, we again illuminate the hologram with another wave, called the reconstruction wave (which in most cases is identical to the reference wave used during the formation of the hologram); this process is termed as reconstruction (see Fig. 18.2). The reconstruction process leads, in general, to a virtual and a real image of the object scene. The virtual image has all the characteristics of the object like parallax, etc. Thus one can move the position of the eye and look behind the objects or one can focus at different distances. The real image can be photographed without the aid of lenses just by placing a light sensitive medium at the

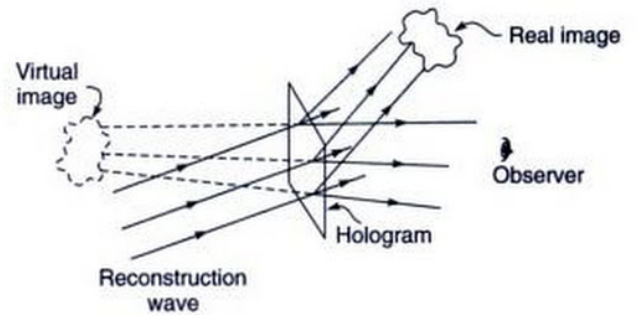


Fig. 18.2 Reconstruction process.

position where the real image is formed. Figures 18.3(a), (b) and (c) represent the object, its hologram and the reconstructed image respectively.

18.2 THEORY

If the object is a point scatterer, then the object wave would just be $\frac{A}{r} \cos(kr - \omega t + \phi)$ where r represents the distance of the point of observation from the point scatterer and A represents a constant; $k = 2\pi/\lambda$. Any general object can be

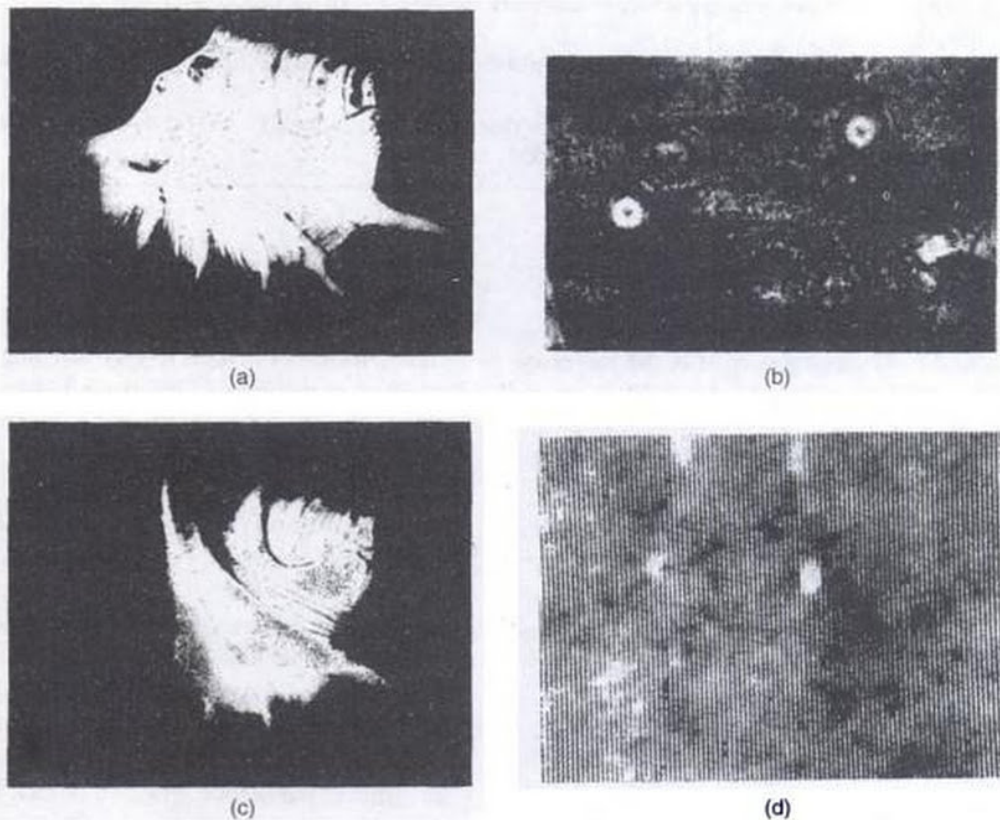


Fig. 18.3 (a) An ordinary photograph of an object. (b) The hologram of the object produced by a method similar to the one as shown in Fig. 18.1. (c) The reconstructed image as seen by an observer. (d) A magnified view of a small portion of the hologram shown in (b). (Photographs courtesy Professor R. S. Sirohi.)

thought of as being made up of a large number of points and the composite wave reflected by the object would be vectorial sum of these. The fundamental problem in holography is the recording of this object wave, in particular, the phase distribution associated with it.

Let us consider the recording process. Let

$$O(x, y) = a(x, y) \cos [\phi(x, y) - \omega t] \quad (1)$$

represents the object wave (which, as mentioned earlier, is due to the superposition of waves from point scatterers on the object) in the plane of the photographic plate which is assumed to be $z = 0$ (see Fig. 18.1). We consider a plane reference wave and assume, for simplicity, that it is propagating in the x - z -plane inclined at an angle θ with the z -direction (see Fig. 18.1). Thus, the field associated with this plane wave would be given by

$$\begin{aligned} r(x, y, z) &= A \cos [\mathbf{k} \cdot \mathbf{r} - \omega t] \\ &= A \cos (kx \sin \theta + kz \cos \theta - \omega t) \end{aligned} \quad (2)$$

If $r(x, y)$ represents the field at the plane $z = 0$ due to this reference wave, then one can see that

$$\begin{aligned} r(x, y) &= A \cos [kx \sin \theta - \omega t] \\ &= A \cos [2\pi\alpha x - \omega t] \end{aligned} \quad (3)$$

where $\alpha = \sin \theta / \lambda$ is the spatial frequency (see Sec. 16.11). The above equation represents the field due to a plane wave inclined at an angle θ with the z -axis and as can be seen the phase varies linearly with x . Notice that there is no y -dependence because the plane wave has been assumed to have its propagation vector in the x - z -plane. Thus the total field at the photographic plate (which is coincident with the plane $z = 0$) would be given by

$$u(x, y, t) = a(x, y) \cos [\phi(x, y) - \omega t] + A \cos [2\pi\alpha x - \omega t] \quad (4)$$

The photographic plate responds only to the intensity which would be proportional to the time average of $[u(x, y, t)]^2$. Thus, the intensity pattern recorded by the photographic plate would be

$$\begin{aligned} I(x, y) &= \langle u^2(x, y, t) \rangle \\ &= \langle [a(x, y) \cos \{\phi(x, y) - \omega t\} \\ &\quad + A \cos(2\pi\alpha x - \omega t)]^2 \rangle \end{aligned} \quad (5)$$

where the angular brackets denote time averaging (see Sec. 15.3). Thus

$$\begin{aligned} I(x, y) &= a^2(x, y) \langle \cos^2 \{\phi(x, y) - \omega t\} \rangle \\ &\quad + A^2 \langle \cos^2 (2\pi\alpha x - \omega t) \rangle \\ &\quad + 2a(x, y) A \langle \cos \{\phi(x, y) - \omega t\} \cos (2\pi\alpha x - \omega t) \rangle \end{aligned} \quad (6)$$

Since

$$\langle \cos^2 [\phi(x, y) - \omega t] \rangle = \frac{1}{2} = \langle \cos^2 (2\pi\alpha x - \omega t) \rangle \quad (7)$$

and

$$\begin{aligned} \langle \cos [\phi(x, y) - \omega t] \cos (2\pi\alpha x - \omega t) \rangle \\ &= \frac{1}{2} \langle \cos [\phi(x, y) + 2\pi\alpha x - 2\omega t] \rangle \\ &= \frac{1}{2} \langle \cos [\phi(x, y) - 2\pi\alpha x] \rangle \\ &= \frac{1}{2} \cos [\phi(x, y) - 2\pi\alpha x] \end{aligned} \quad (8)$$

Eq. (6) becomes

$$\begin{aligned} I(x, y) &= \frac{1}{2} a^2(x, y) + \frac{1}{2} A^2 \\ &\quad + A a(x, y) \cos [\phi(x, y) - 2\pi\alpha x] \end{aligned} \quad (9)$$

From the above relation it is obvious that the phase information of the object wave, which is contained in $\phi(x, y)$, is recorded in the intensity pattern.

When the photographic plate (which has recorded the above intensity pattern) is developed, one obtains a hologram [see Figs. 18.3(b) and (d)]. The transmittance of the hologram, i.e., the ratio of the transmitted field to the incident field, depends on $I(x, y)$. By a suitable developing process one can obtain a condition under which the amplitude transmittance would be linearly related to $I(x, y)$. Thus, in such a case if $R(x, y)$ represents the field of the reconstruction wave at the hologram plane, then the transmitted field would be given by

$$\begin{aligned} v(x, y) &= K R(x, y) I(x, y) \\ &= K \left[\frac{1}{2} a^2(x, y) + \frac{1}{2} A^2 \right] R(x, y) \\ &\quad + K A a(x, y) R(x, y) \cos [\phi(x, y) - 2\pi\alpha x] \end{aligned} \quad (10)$$

where K is a constant. We consider the case when the reconstruction wave is identical to the reference wave $r(x, y)$ (see Fig. 18.2). In such a case we would obtain (omitting the constant K)

$$\begin{aligned} v(x, y) &= \left[\frac{1}{2} a^2(x, y) + \frac{1}{2} A^2 \right] A \cos (2\pi\alpha x - \omega t) \\ &\quad + A^2 a(x, y) \cos [2\pi\alpha x - \omega t] \cos [\phi(x, y) - 2\pi\alpha x] \\ &= \left[\frac{1}{2} a^2(x, y) + \frac{1}{2} A^2 \right] A \cos (2\pi\alpha x - \omega t) \\ &\quad + \frac{1}{2} A^2 a(x, y) \cos [\phi(x, y) - \omega t] \\ &\quad + \frac{1}{2} A^2 a(x, y) \cos [4\pi\alpha x - \phi(x, y) - \omega t] \end{aligned} \quad (11)$$

Equation (11) gives the transmitted field in the plane $z = 0$. We consider each of the three terms separately. The first term is nothing but the reconstruction wave itself whose amplitude is modulated due to the presence of the term $a^2(x, y)$. This part of the total field is traveling in the direction of the reconstructed wave. The second term is identical (within a constant term) to the RHS of Eq. (1) and hence represents the original object wave; this gives rise to a virtual image. Thus the effect of viewing this wave is the same as viewing the object itself. The reconstructed object wave is traveling in the same direction as the original object wave.

To study the last term we first observe that in addition to the term $4\pi\alpha x$, the phase term $\phi(x, y)$ carries a negative sign. The negative sign represents the fact that the wave has a curvature opposite to that of the object wave. Thus if the object wave is a diverging spherical wave then the last term represents a converging spherical wave. Thus in contrast to the second term, this wave forms a real image of the object which can be photographed by simply placing a film (see Fig. 18.2).

To determine the effect of the term $4\pi\alpha x$, we consider the case when the object wave is also a plane wave traveling along the z -axis. For such a wave $\phi(x, y) = 0$ and the last term would represent a plane wave propagating along a direction $\theta' = \sin^{-1}(2 \sin \theta)$. Thus the effect of the term $4\pi\alpha x$ is to rotate the direction of the wave. Hence the last term on the RHS of Eq. (11) represents the conjugate of the object wave propagating along a direction different from that of the reconstruction wave and the object wave, which forms a real image of the object. Since the waves represented by the three terms are propagating along different directions they separate after traversing a distance and enable the observer to view the virtual image without any disturbance.

A very interesting property possessed by holograms is that even if the hologram is broken up into different fragments, each separate fragment is capable of producing a complete virtual image of the object.* This property can be understood from the fact that for a diffusely reflecting object, each point of the object illuminates the complete hologram and consequently each point in the hologram receives waves from the complete object. But the resolution in the image decreases as the size of the fragment decreases. For non-diffusely reflecting objects or for transparencies, one makes use of an additional diffusing screen through which the object is illuminated.

Example 18.1 As an explicit example of the formation and reconstruction of a hologram, we consider the simple case when both the object wave and the reference wave

are plane waves [see Fig. 18.4(a)]—a plane object wave corresponds to a single object point lying far away from the hologram. (a) Show that for such a case, the hologram consists of a series of Young's interference fringes having an intensity distribution of the \cos^2 type. (see also Fig. 12.11) (b) If we reconstruct the hologram with another plane wave [see Fig. 18.4(b)], then show that the transmitted light consists of a zero-order plane wave and two first-order plane waves; the two first-order waves correspond to the primary and conjugate waves.

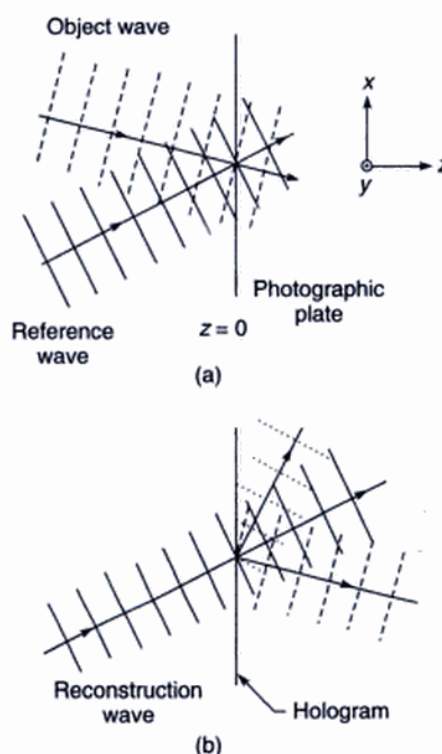


Fig. 18.4 (a) Formation of a hologram, when both the object wave and the reference wave are plane waves. (b) Reconstruction of the hologram with another plane wave.

Solution: (a) Consider a plane wave with its propagation vector lying in the x - z -plane and making an angle θ_1 with the z -axis. For such a wave, the field is of the form

$$A_1 \cos [kx \sin \theta_1 + kz \cos \theta_1 - \omega t]$$

If the photographic film is assumed to coincide with the plane $z = 0$, then the field distribution on this plane would be given by

$$A_1 \cos [kx \sin \theta_1 - \omega t]$$

*This property of a hologram exists only when the object is a diffuse scatterer such that the wave from each scattering point of the object reaches all parts of the hologram plate. There are cases where this does not hold good; for example, when a hologram of a transparency is to be recorded.

Similarly the field (on the plane of the film) due to a plane wave making an angle θ_2 with the z -axis, will be given by

$$A_2 \cos [kx \sin \theta_2 - \omega t]$$

The resultant intensity distribution would be proportional to

$$\begin{aligned} & \langle [A_1 \cos \{kx \sin \theta_1 - \omega t\} + A_2 \cos \{kx \sin \theta_2 - \omega t\}]^2 \rangle \\ &= \frac{1}{2} A_1^2 + \frac{1}{2} A_2^2 + A_1 A_2 \cos [kx (\sin \theta_1 - \sin \theta_2)] \\ &= \frac{1}{2} (A_1 - A_2)^2 + 2A_1 A_2 \cos^2 \left[\frac{kx}{2} (\sin \theta_1 - \sin \theta_2) \right] \end{aligned}$$

For $A_1 = A_2$, the above expression simplifies to

$$2A^2 \cos^2 \left[\frac{kx}{2} (\sin \theta_1 - \sin \theta_2) \right]$$

showing that the intensity remains constant along lines parallel to the y -axis with fringe spacing depending on the values of θ_1 and θ_2 . Further, the intensity distribution is of the \cos^2 type (cf. Fig. 12.11).

(b) Before we calculate the transmitted field of the hologram, we first consider a narrow slit of width b being illuminated by a plane wave (see Fig. 18.5). Consider an element ds at a distance s from the center of the slit. Then the amplitude at a far away point P due to this element would be proportional to $\sin [k(r - s \sin \theta) - \omega t] ds$; here $k = 2\pi/\lambda$ and θ is defined in Fig. 18.5. Thus the total field in the direction θ would be given by

$$E = A \int_{-b/2}^{+b/2} \sin [k(r - s \sin \theta) - \omega t] ds \quad (12)$$

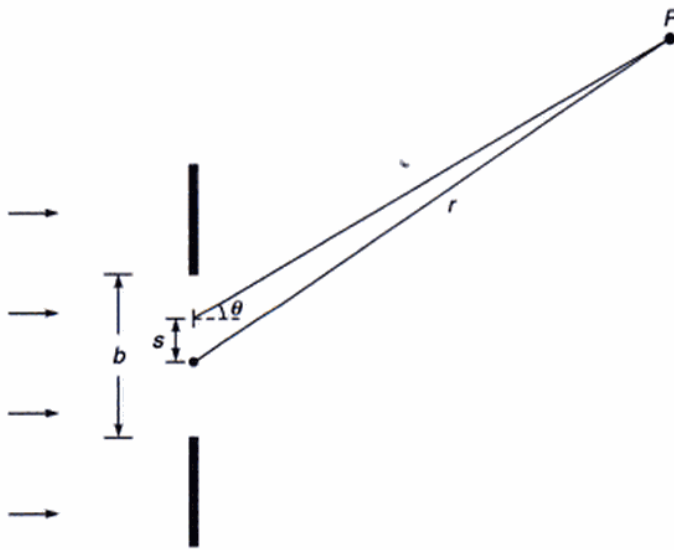


Fig. 18.5 A plane wave incident on a narrow slit of width b .

where A is a constant. The above integral can also be written as

$$\begin{aligned} E &= A \int_{-b/2}^{+b/2} [\sin (kr - \omega t) \cos (ks \sin \theta) \\ &\quad - \cos (kr - \omega t) \sin (ks \sin \theta)] ds \\ &= 2A \sin (kr - \omega t) \frac{\sin \left(\frac{kb}{2} \sin \theta \right)}{k \sin \theta} \end{aligned}$$

where the second integral is zero because the integrand is an odd function of s . Thus

$$E = Ab \sin (kr - \omega t) \frac{\sin \beta}{\beta} \quad (13)$$

where

$$\beta = \frac{1}{2} kb \sin \theta = \frac{\pi b \sin \theta}{\lambda}$$

which is of the same form as obtained in Sec. 16.2. In the present case, the hologram has a $\cos^2 \alpha s$ type of variation in transmittance and hence the transmitted field will be of the form

$$E = A \int_{-b/2}^{+b/2} \cos^2 \alpha s \sin [kr - ks (\sin \theta - \sin \theta_i) - \omega t] ds \quad (14)$$

where θ_i represents the angle of incidence of the illuminating plane wave. Thus

$$\begin{aligned} E &= \frac{1}{2} \int_{-b/2}^{+b/2} (1 + \cos 2\alpha s) \\ &\quad [\sin (kr - \omega t) \cos \{ks (\sin \theta - \sin \theta_i)\} \\ &\quad - \cos (kr - \omega t) \sin \{ks (\sin \theta - \sin \theta_i)\}] ds \\ &= \frac{1}{2} A \sin (kr - \omega t) \left[\int_{-b/2}^{+b/2} \cos \{ks (\sin \theta - \sin \theta_i)\} ds \right. \\ &\quad + \frac{1}{2} \int_{-b/2}^{+b/2} \cos \{ks (\sin \theta - \sin \theta_i + 2\alpha)\} ds \\ &\quad \left. + \frac{1}{2} \int_{-b/2}^{+b/2} \cos \{ks (\sin \theta - \sin \theta_i - 2\alpha)\} ds \right] \quad (15) \end{aligned}$$

The above integrations can easily be carried out. Thus, for example,

$$\int_{-b/2}^{+b/2} \cos \{ks (\sin \theta - \sin \theta_i + 2\alpha)\} ds = \frac{\sin \left[b \frac{k}{2} (\sin \theta - \sin \theta_i + 2\alpha) \right]}{\frac{k}{2} (\sin \theta - \sin \theta_i + 2\alpha)} \quad (16)$$

which becomes more and more sharply peaked around $\sin \theta = \sin \theta_i - 2\alpha$ as $b \rightarrow \infty$, i.e., as the size of the hologram becomes larger. Thus the three integrals in Eq. (15) in the limit of a large value of b give rise to three plane waves propagating along $\sin \theta = \sin \theta_i$, $\sin \theta = \sin \theta_i - 2\alpha$ and $\sin \theta = \sin \theta_i + 2\alpha$, which represent the zero-order and two first order waves.

Example 18.2 Consider the formation of a hologram with a point object and a plane reference wave (see Fig. 12.13(a)). Choose the z -axis to be along the normal from the point source to the plane of the photograph, assumed to be coincident with the plane $z = 0$. For simplicity assume the reference wave to fall normally on the photographic plate. Obtain the interference pattern recorded by the hologram.

Solution: Let the point source be situated at a distance d from the photographic plate. The field at any point $P(x, y, 0)$ on the photographic plate, due to waves emanating from the point object would be given by

$$O(x, y, z = 0, t) = \frac{A}{r} \cos(kr - \omega t) \quad (17)$$

where $r = (x^2 + y^2 + d^2)^{1/2}$ and A represents a constant. A plane wave traveling along a direction parallel to the z -axis would be given by

$$R(x, y, z, t) = B \cos(kz - \omega t) \quad (18)$$

Hence, the field due to the reference wave at the plane of the photographic plate ($z = 0$) would be

$$R(x, y, z = 0, t) = B \cos \omega t \quad (19)$$

Thus, the total field at the plane of the photographic plate would be

$$\begin{aligned} T(x, y, t) &= O(x, y, z = 0, t) + R(x, y, z = 0, t) \\ &= \frac{A}{r} \cos(kr - \omega t) + B \cos \omega t \end{aligned} \quad (20)$$

The recorded intensity pattern would be

$$\begin{aligned} I(x, y) &= \langle |T(x, y, t)|^2 \rangle \\ &= \left\langle \left| \frac{A}{r} \cos(kr - \omega t) + B \cos \omega t \right|^2 \right\rangle \end{aligned} \quad (21)$$

where, as before, angular brackets denote time averaging. Carrying out the above time averaging, we get

$$I(x, y) = \frac{A^2}{2r^2} + \frac{B^2}{2} + \frac{AB}{r} \cos kr \quad (22)$$

If we assume that $d \gg x, y$ (which is valid in most practical cases), we can write

$$r = (x^2 + y^2 + d^2)^{1/2} \approx d + \frac{x^2 + y^2}{2d} \quad (23)$$

Thus

$$I(x, y) = \frac{A^2}{2d^2} + \frac{B^2}{2} + \frac{AB}{r} \cos \left[kd + \frac{k}{2d} (x^2 + y^2) \right] \quad (24)$$

The resultant fringe pattern is circular and centered at the origin (see Example 12.7). The hologram thus formed is essentially a zone plate with the transmittance varying sinusoidally in contrast to the Fresnel zone plate [see Fig. 12.13(b) and Sec. 17.3].

18.3 REQUIREMENTS

Since holography is essentially an interference phenomenon, certain coherence requirements have to be met with. In Chapter 15 we had introduced the notion of coherence length. Thus, if stable interference fringes are to be formed (so that they are recordable), the maximum path difference between the object wave and the reference wave should not exceed the coherence length. Further, the spatial coherence is important so that the waves scattered from different regions of the object could interfere with the reference beam.

During reconstruction, the reconstructed image depends both on the wavelength and the position of the reconstructing source. Hence if the resolution in the reconstructed image has to be good, the source must not be broad and must be emitting a narrow band of wavelengths. It may be worthwhile mentioning here that the reconstruction process has associated with it aberrations similar to that in the images formed by lenses. If the reconstruction source is of the same wavelength and is situated at the same relative position with respect to the hologram as the reference source, then the reconstructed image does not suffer from any aberrations.

Thus, whenever

$$\phi'(x, y) - \phi(x, y) = 2m\pi, m = 0, 1, 2, \dots \quad (28)$$

the two waves would interfere constructively and whenever,

$$\phi'(x, y) - \phi(x, y) = (2m + 1) \frac{\pi}{2}; m = 0, 1, 2, \dots \quad (29)$$

the two waves interfere destructively. Thus, depending on $[\phi'(x, y) - \phi(x, y)]$, one obtains, on reconstruction, the object superimposed with bright and dark fringes (see Fig. 18.7).

We will consider here a simple application of the above technique in the determination of the Young's modulus of a material. If we have a bar fixed at one end and loaded at the other and if it results in a displacement δ of the end of the bar, then one can show that*

$$\delta = \frac{WL^3}{3YI} \quad (30)$$

where W is the load, L is the length of the bar, I is the moment of inertia of cross-section which for a rectangular bar of width a and thickness b , is given by $I = ab^3/12$; Y represents the Young's modulus of the material of the rod. Thus if we could determine δ for a given load, then Y can be determined from Eq. (30).

We will first determine an expression for $(\phi' - \phi)$. In Fig. 18.6 we have shown the undisplaced and displaced positions of the cantilever illuminated by a laser light along a direction making an angle θ_1 with the z -axis. We observe the cantilever along a direction making an angle θ_2 with the z -axis. The phase change when the cantilever undergoes a displacement δ as shown in Fig. 18.6(b) would be

$$\begin{aligned} \phi' - \phi &= \frac{2\pi}{\lambda} (\delta \cos \theta_1 + \delta \cos \theta_2) \\ &= \frac{2\pi}{\lambda} \delta (\cos \theta_1 + \cos \theta_2) \end{aligned} \quad (31)$$

If there are N fringes over the length L of the cantilever, then since a phase difference of 2π corresponds to one fringe [see Eq. (28)] we can write

$$\frac{2\pi}{\lambda} \delta (\cos \theta_1 + \cos \theta_2) = N \cdot 2\pi$$

or

$$\delta = \frac{N\lambda}{(\cos \theta_1 + \cos \theta_2)}$$

Thus by measuring N , θ_1 and θ_2 and knowing λ , δ can be determined. Figure 18.7 shows the reconstruction of a double exposed hologram of an aluminum strip of width 4 cm, thickness 0.2 cm and of length 12 cm. From the number of fringes formed, one can calculate the Young's modulus (see Problem 18.3).

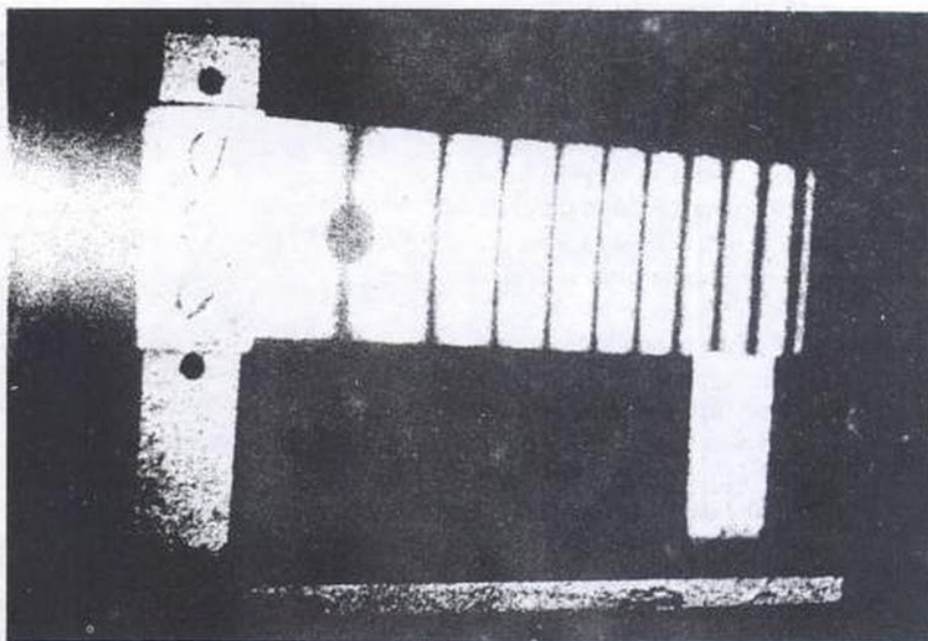


Fig. 18.7 Interference fringes produced in the measurement of Young's modulus using double exposure interferometry. (Photograph courtesy Professor R. S. Sirohi.)

*See, e.g., Ref. 13, p.75.

SUMMARY

- The basic technique in holography is the following : In the recording of the hologram, one superimposes on the object wave another wave called the reference wave and the photographic plate is made to record the resulting interference pattern. The reference wave is usually a plane wave. This recorded interference pattern forms the hologram and contains information not only about the amplitude but also about the phase of the object wave. To view the image, we again illuminate the hologram with another wave, called the reconstruction wave. The reconstruction process leads, in general, to a virtual and a real image of the object scene. The virtual image has all the characteristics of the object like parallax, etc.
- If the object wave and the reference wave are plane waves, the hologram consists of a series of Young's interference fringes.
- For a point object and a plane reference wave, the hologram is very similar to a zone plate with the transmittance varying sinusoidally in contrast to the Fresnel zone plate.

PROBLEMS

- 18.1** Consider the reconstruction of the hologram as formed in the configuration of Example 18.2, by a plane wave traveling along a direction parallel to the z -axis. Show the formation of a virtual and a real image.
- 18.2** In continuation of Example 18.2, calculate the interference pattern when the incident plane wave makes an angle θ with the z -axis [see Fig. 12.13]. Assume $B = A/d$.

$$\left[\text{Ans. } 4B^2 \cos^2 \left\{ kd - kx \sin \theta + \frac{k}{2d}(x^2 + y^2) \right\} \right]$$

- 18.3** Figure 18.7 corresponds to the reconstruction of a doubly exposed hologram, the objects corresponding to the unstrained and strained positions of an aluminum bar of width 4 cm, thickness 0.2 cm and length 12 cm. If the strained position corresponds to a load of 1 gm force applied at the end of the bar, calculate the Young's modulus of aluminum. Assume $\theta_1 = \theta_2 = 0$; assume $\lambda = 6328 \text{ \AA}$.

[Hint: N represents the number of fringes produced over the length of the cantilever.]

$$[\text{Ans. } 0.7 \times 10^{11} \text{ N/m}^2]$$

REFERENCES AND SUGGESTED READINGS

1. D. Gabor, 'A New Microscopic Principle', *Nature*, Vol. **161**, 777, 1948; 'Microscopy by Reconstructed Wavefronts', *Proceedings of the Royal Society (London)*, Vol. **A197**, 454, 1949.
2. K. Thyagarajan and A. K. Ghatak, *Lasers: Theory and Applications*, Plenum Press, New York, 1981 (Reprinted by Macmillan India Ltd., New Delhi.)
3. J. C. Brown and J. A. Harte, 'Holography in the undergraduate optics course', *American Journal of Physics*, Vol. **37**, 441, 1969.
4. H. J. Caulfield and S. Lu, *The Applications of Holography*, John Wiley & Sons, New York, 1970.
5. R. J. Collier, C. B. Burckhardt and L. H. Lin, *Optical Holography*, Academic Press, New York, 1971.
6. A. K. Ghatak and K. Thyagarajan, *Contemporary Optics*, Plenum Press, New York, 1978 (Reprinted by Macmillan, New Delhi, 1984.)
7. M. P. Givens, 'Introduction to holography', *American Journal of Physics*, Vol. **35**, 1056, 1967.
8. E. N. Leith and J. Upatnieks, 'Photography by Laser', *Scientific American*, Vol. **212**, p. 24, June, 1965.
9. A. F. Methernal, 'Acoustical holography', *Scientific American*, Vol. **221**, p. 36, October, 1969.
10. K. S. Pennington, 'Advances in holography', *Scientific American*, Vol. **218**, p. 40, February, 1968.
11. H. M. Smith, *Principles of Holography*, Wiley Interscience, New York, 1975.
12. D. Venkateshwarulu, 'Holography, theory and applications', *Journal of Scientific and Industrial Research*, Vol. **29**, November, 1970.
13. B. L. Worsnop and H. T. Flint, *Advanced Practical Physics for Students*, Asia Publishing House, Bombay, 1951.

PART 5

Electromagnetic Character of Light

This part consists of three chapters discussing various aspects of the electromagnetic character of light waves. In the first chapter of this part, the generation and analysis of various forms of polarized light is discussed followed by a detailed analysis of propagation of electromagnetic waves in anisotropic media including first principle derivations of wave and ray velocities. Applications like optical activity, Faraday rotation etc. have also been discussed. Chapter 20 is a bit mathematical – starting with Maxwell's equations, various states of polarization are discussed; the wave equation has also been derived which had led Maxwell to predict the existence of electromagnetic waves. Reflection and refraction of electromagnetic waves by a dielectric interface has been discussed in chapter 21. The results directly explain phenomena like Brewster's law, total internal reflection, evanescent waves, Fabry-Perot transmission resonances, etc.

Chapter 19

Polarization and Double Refraction

As to the other emanation which should produce the irregular refraction, I wished to try what Elliptical waves, or rather spheroidal waves, would do; and these I suppose would spread differently both in the ethereal matter diffused throughout the crystal and in the particles of which it is composed...

Christiaan Huygens

Important Milestones

1669	Erasmus Bartholinus discovered double refraction in calcite
1678	In the wave theory of light communicated to the Academie des Science in Paris, Christiaan Huygens gave the theory of double refraction in calcite discovered by Bartholinus
1815	David Brewster showed polarization of light by reflection
1828	William Nicol invented the prism which produced polarized light – this prism came to be known as the Nicol Prism.

19.1 INTRODUCTION

If we move one end of a string up and down then a transverse wave is generated [see Fig. 19.1(a)]. Each point of the string executes a sinusoidal oscillation in a straight line (along the x -axis) and the wave is, therefore, known as a *linearly polarized wave*. It is also known as a plane polarized wave because the string is always confined to the x - z plane. The displacement for such a wave can be written in the form

$$\text{and} \quad \left. \begin{aligned} x(z, t) &= a \cos(kz - \omega t + \phi_1) \\ y(z, t) &= 0 \end{aligned} \right\} \quad (1)$$

where a represents the amplitude of the wave and ϕ_1 is the phase constant to be determined from the boundary condition; the y -coordinate of the displacement is always zero. At any instant the displacement will be a cosine curve as shown in Fig. 19.1(a). Further, an arbitrary point $z = z_0$ will execute a simple harmonic motion of amplitude a . The string can also be made to vibrate in the y - z plane [see Fig. 19.1(b)] for which the displacement would be given by

$$\text{and} \quad \left. \begin{aligned} x(z, t) &= a \cos(kz - \omega t + \phi_2) \\ y(z, t) &= 0 \end{aligned} \right\} \quad (2)$$

In general, the string can be made to vibrate in any plane containing the z -axis. If one rotates the end of the string on the circumference of a circle then each point of the string will move in a circular path as shown in Fig. 19.2; such a wave is known as a circularly polarized wave and the corresponding displacement would be given by

$$\text{and} \quad \left. \begin{aligned} x(z, t) &= a \cos(kz - \omega t + \phi) \\ y(z, t) &= a \sin(kz - \omega t + \phi) \end{aligned} \right\} \quad (3)$$

so that $x^2 + y^2$ is a constant ($= a^2$).

We next consider a long narrow slit placed in the path of the string as shown in Fig. 19.3(a). If the length of the slit is along the direction of the displacement then the entire amplitude will be transmitted as shown in Fig. 19.3(a). On the other hand, if the slit is at right angles to the direction of the displacement then almost nothing will be transmitted to the other side of the slit [see Fig. 19.3(b)]. This is because of the fact that the slit allows only the component of the displacement, which is along the length of the slit, to pass through. We must mention here that if a longitudinal wave was propagating through the string then the amplitude of the transmitted wave would have been the same for all orientations of the slit. Thus, the change in the amplitude of the transmitted wave with the orientation of the slit is due to the transverse character of the wave. Indeed, an experiment

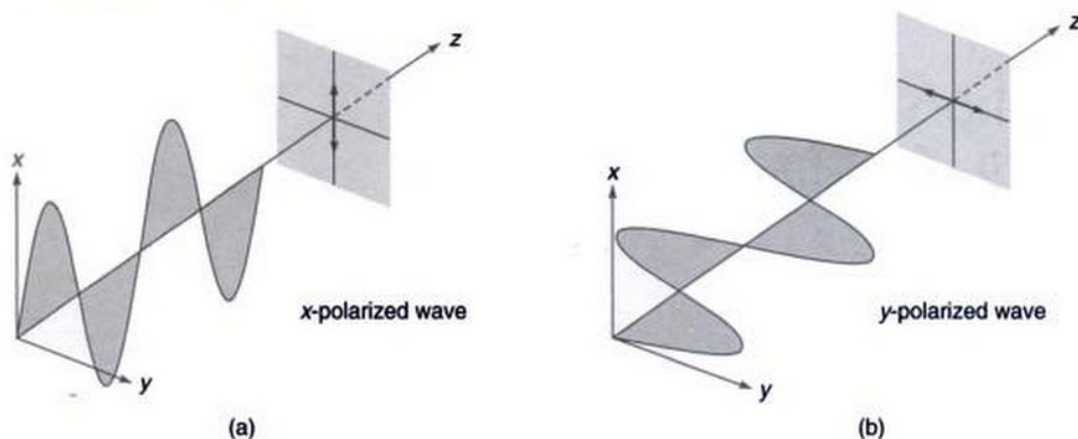


Fig. 19.1 (a) A linearly polarized wave on a string with the displacement confined to the x - z plane; (b) A linearly polarized wave on a string with the displacement confined to the y - z plane.

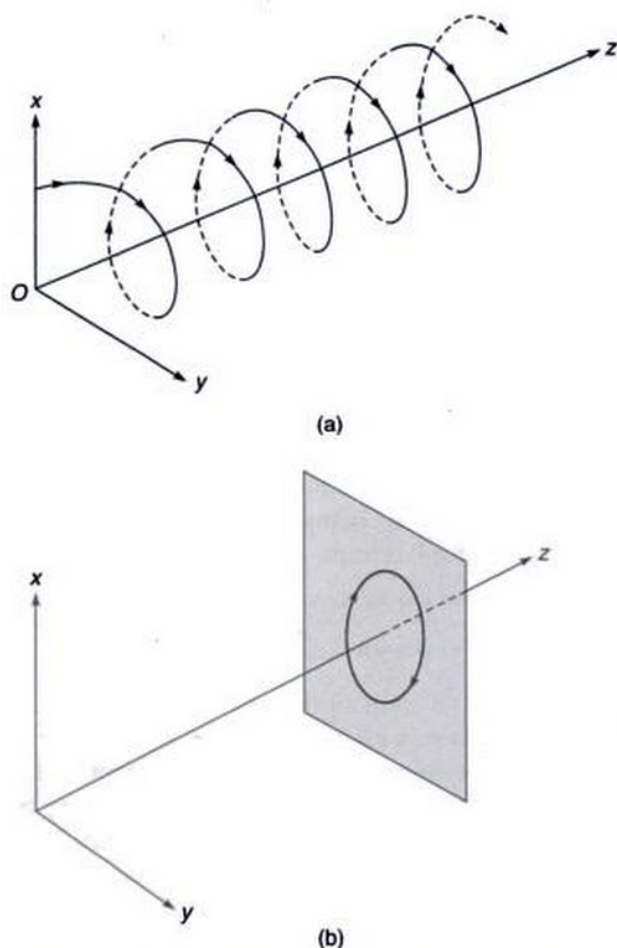


Fig. 19.2 (a) The displacement corresponding to a circularly polarized wave – all points on the string are at the same distance from the z -axis. (b) Each point on the string rotates on the circumference of the circle.

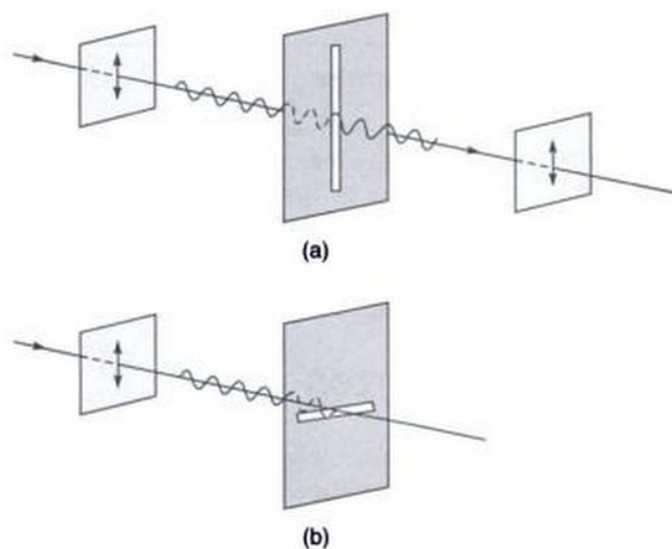


Fig. 19.3 If a linearly polarized transverse wave (propagating on a string) is incident on a long narrow slit, then the slit will allow only the component of the displacement, which is along the length of the slit, to pass through.

which is, in principle, very similar to the experiment discussed above proves the transverse character of light waves. However, before we discuss the experiment with light waves we must define an unpolarized wave.

We once again consider transverse waves generated at one end of a string. If the plane of vibration is changed in a random manner in very short intervals of time, then such a wave is known as an unpolarized wave*. If an unpolarized wave falls on a slit S_1 (see Fig. 19.4) then the displacement associated with the transmitted wave will be along the

*By a short interval, we imply times which are short compared to the detection time; however, for the wave to be characterized with a certain frequency ν , this time has to be much greater than $1/\nu$, so that in the short interval it executes a large number of oscillations (see also Sec. 15.1).

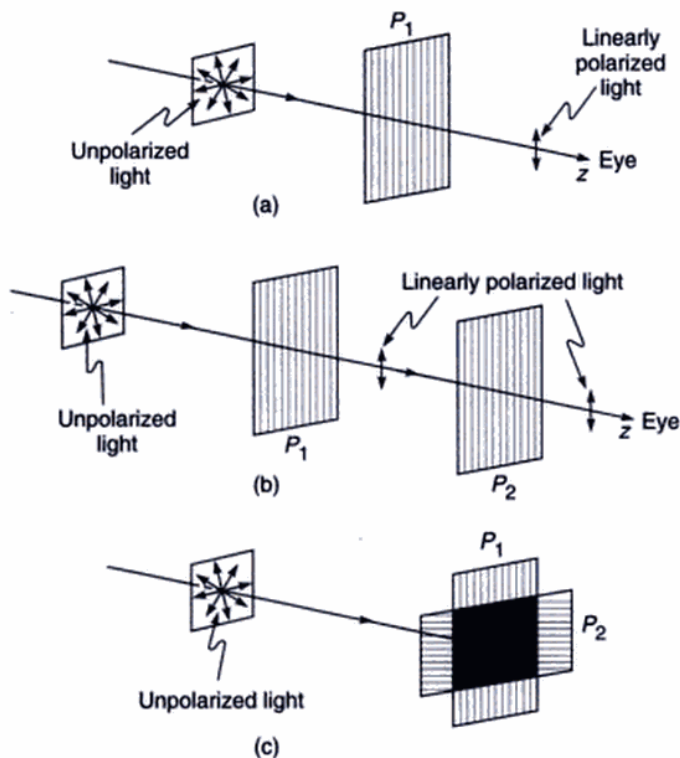


Fig. 19.6 If an ordinary light beam is allowed to fall on a Polaroid, then the emerging beam will be linearly polarized and if we place another Polaroid P_2 , then the intensity of the transmitted light will depend on the relative orientation of P_2 with respect to P_1 .

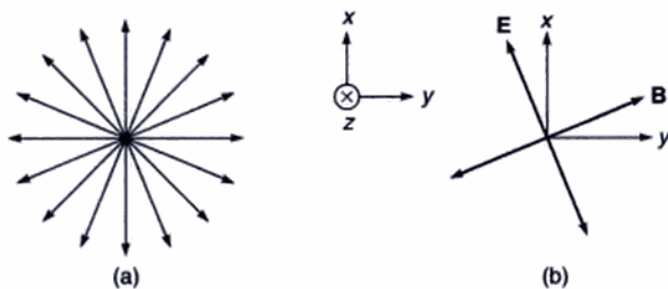


Fig. 19.7 (a) For an unpolarized wave propagating in the $+z$ -direction, the electric vector (which lies in the x - y plane) continues to change its direction in a random manner. (b) For a linearly polarized wave, the electric (or the magnetic) vector oscillates along a particular direction.

discussions, this phenomenon proves the transverse character of light. i.e., the displacement associated with a light wave is at right angles to the direction of propagation

*For further details, see Ref. 1.

of the wave. The polaroid P_1 acts as a polarizer and the transmitted beam is linearly polarized. The second polaroid acts as an analyzer.

19.2 PRODUCTION OF POLARIZED LIGHT

In this section we will discuss various methods for the production of linearly polarized light waves.

19.2.1 The Wire Grid Polarizer and the Polaroid

The physics behind the working of the wire grid polarizer is probably the easiest to understand. It essentially consists of a large number of thin copper wires placed parallel to each other as shown in Fig. 19.8. When an unpolarized electromagnetic wave is incident on it then the component of the electric vector along the length of the wire is absorbed. This is due to the fact that the electric field does work on the electrons inside the thin wires and the energy associated with the electric field is lost in the Joule heating of the wires. On the other hand, (since the wires are assumed to be very thin) the component of the electric vector along the x -axis passes through without much attenuation. Thus the emergent wave is linearly polarized with the electric vector along the x -axis. However, for the system to be effective (i.e., for the E_y component to be almost completely attenuated) the spacing between the wires should be $\leq \lambda$. Clearly, the fabrication of such a polarizer for a 3 cm microwave is relatively easy because the spacing has to be ≤ 3 cm. On the other hand, since the light waves are associated with a very small wavelength ($\sim 5 \times 10^{-5}$ cm), the fabrication of a polarizer in which the wires are placed at distances $\leq 5 \times 10^{-5}$ cm is extremely difficult. Nevertheless, Bird and Parrish did succeed in putting about 30,000 wires in about one inch.* The details of the procedure for making this wire

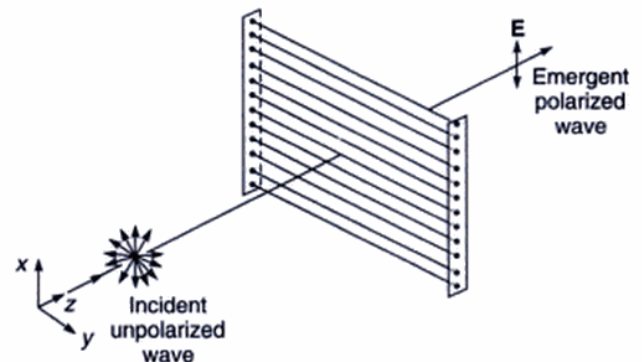


Fig. 19.8 The wire-grid polarizer.

A simple method for eliminating one of the beams is through selective absorption; this property of selective absorption is known as dichroism. A crystal like tourmaline has different coefficients of absorption for the two linearly polarized beams into which the incident beam splits up. Consequently, one of the beams gets absorbed quickly and the other component passes through without much attenuation. Thus, if an unpolarized beam is passed through a tourmaline crystal, the emergent beam will be linearly polarized (see Fig. 19.10).

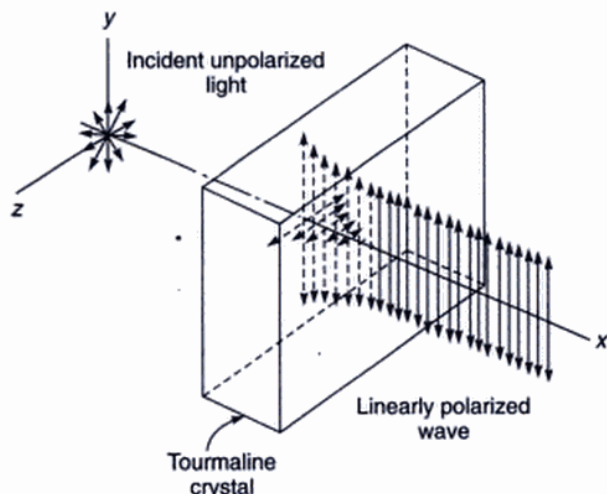


Fig. 19.10 When an unpolarized beam enters a dichroic crystal like tourmaline, it splits up into two linearly polarized components. One of the components gets absorbed quickly and the other component passes through without much attenuation. [Adapted from Ref. 3; used with permission.]

Another method for eliminating one of the polarized beams is through total internal reflection. We will show in Secs 19.5 and 19.12 that the two beams have different ray velocities and as such the corresponding refractive indices will be different. If one can sandwich a layer of a material whose refractive index lies between the two, then for one of the beams, the incidence will be at a rarer medium and for the other it will be at a denser medium. This principle is used in a Nicol prism which consists of a calcite crystal cut in such a way that for the beam, for which the sandwiched material is a rarer medium, the angle of incidence is greater than the critical angle. Thus this particular beam will be eliminated by total internal reflection. Figure 19.11 shows a properly cut calcite crystal in which a layer of Canada Balsam has been introduced so that the ordinary ray undergoes total internal reflection. The extra-ordinary component passes through and the beam emerging from the crystal is linearly polarized.

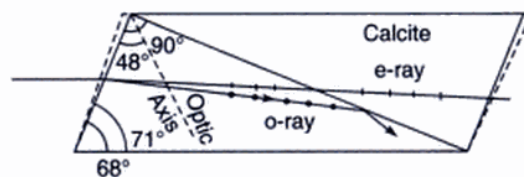


Fig. 19.11 The Nicol prism. The dashed outline corresponds to the natural crystal which is cut in such a way that the ordinary ray undergoes total internal reflection at the Canada Balsam layer.

19.2.4 Polarization by Scattering

If an unpolarized beam is allowed to fall on a gas, then the beam scattered at 90° to the incident beam is linearly polarized. This follows from the fact that the waves propagating in the y -direction are produced by the x -component of the dipole oscillations (see Fig. 19.12). The y -component of the dipole oscillations will produce no field in the y -direction (see Sec. 20.4). It may be of interest to mention that it was through scattering experiments that Barkla could establish the transverse character of X-rays. Clearly, if the incident beam is linearly polarized with its electric vector along the x -direction, then there will be no scattered light along the x -axis. As such, one can carry out an analysis of a scattered wave by allowing it to undergo a further scattering [see Fig. 19.12(b)].

19.3 MALUS' LAW

Let us consider a polarizer P_1 which has a pass-axis parallel to the x -axis (see Fig. 19.13); i.e., if an unpolarized beam propagating in the z -direction is incident on the polarizer, then the electric vector associated with the emergent wave will oscillate along the x -axis. It should be noted that if the polarizer is a polaroid, then for the pass axis to be along the x -direction, the long chain molecules must be aligned along the y -axis. We next consider the incidence of the x -polarized beam on the polaroid P_2 whose pass axis makes an angle θ with the x -axis (see Fig. 19.13). If the amplitude of the incident electric field is E_0 , then the amplitude of the wave emerging from the polaroid P_2 will be $E_0 \cos \theta$ and thus the intensity of the emerging beam will be given by

$$I = I_0 \cos^2 \theta \quad (10)$$

where I_0 represents the intensity of the emergent beam when the pass axis of P_2 is also along the x -axis (i.e., when $\theta = 0$). Equation (10) represents the Malus' Law. Thus, if a linearly polarized beam is incident on a polaroid and if the polaroid is rotated about the z -axis, then the intensity of the emergent wave will vary according to the above law. For example, if

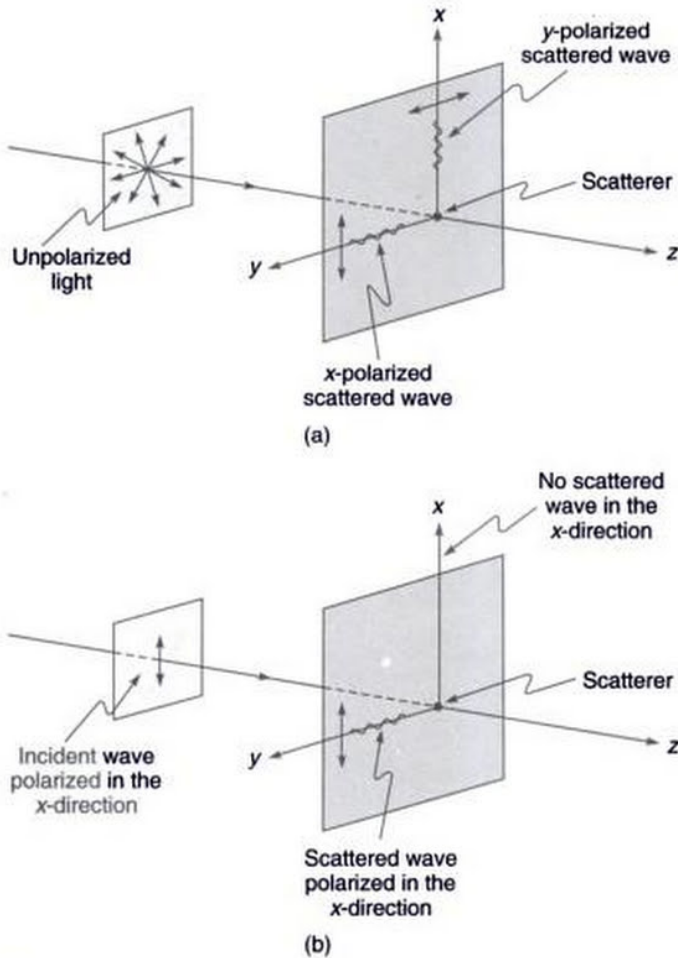


Fig. 19.12 (a) If the electromagnetic wave is propagating along the z -direction, then the scattered wave along any direction is perpendicular to the z -axis will be linearly polarized. (b) If a linearly polarized wave (with its \mathbf{E} oscillating along the x -direction) is incident on a dipole, then there will be no scattered wave in the x -direction.

the polaroid P_2 shown in Fig. 19.13 is rotated in the clockwise direction, then the intensity will increase till the pass-axis is parallel to the x -axis; a further rotation will result in a decrease in intensity till the pass-axis is parallel to the y -axis, where the intensity will be almost zero. If we further rotate it, it will pass through a maximum and again a minimum before it reaches its original position.

19.4 SUPERPOSITION OF TWO DISTURBANCES

Let us consider the propagation of two linearly polarized electromagnetic waves (both propagating along the z -axis)

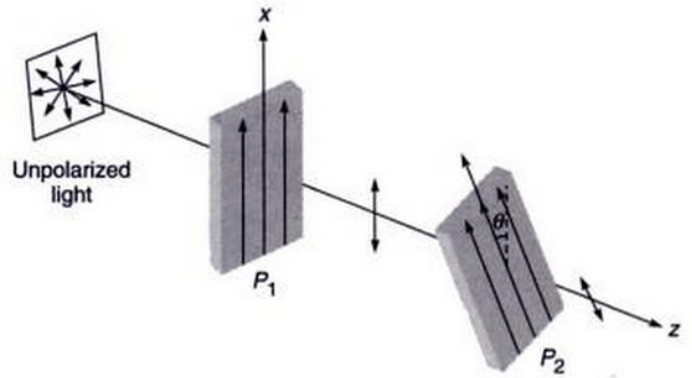


Fig. 19.13 An unpolarized light beam gets x -polarized after passing through the polaroid P_1 , the pass axis of the second polaroid P_2 makes an angle θ with the x -axis. The intensity of the emerging beam will vary as $\cos^2 \theta$.

with their electric vectors oscillating along the x -axis. The electric fields associated with the waves can be written in the form

$$\mathbf{E}_1 = \hat{\mathbf{x}} a_1 \cos(kz - \omega t + \theta_1) \quad (11)$$

$$\mathbf{E}_2 = \hat{\mathbf{x}} a_2 \cos(kz - \omega t + \theta_2) \quad (12)$$

where a_1 and a_2 represent the amplitudes of the waves; $\hat{\mathbf{x}}$ represents the unit vector along the x -axis and θ_1 and θ_2 are phase constants. The resultant of these two waves would be given by

$$\mathbf{E} = \mathbf{E}_1 + \mathbf{E}_2 \quad (13)$$

which can always be written in the form

$$\mathbf{E} = \hat{\mathbf{x}} a \cos(kz - \omega t + \theta) \quad (14)$$

where

$$a = [a_1^2 + a_2^2 + 2a_1 a_2 \cos(\theta_1 - \theta_2)]^{1/2} \quad (15)$$

represents the amplitude of the wave. Equation (14) tells us that the resultant is also a linearly polarized wave with its electric vector oscillating along the same axis.

We next consider the superposition of two linearly polarized electromagnetic waves (both propagating along the z -axis) but with their electric vectors oscillating along two mutually perpendicular directions. Thus, we may have

$$\mathbf{E}_1 = \hat{\mathbf{x}} a_1 \cos(kz - \omega t) \quad (16)$$

$$\mathbf{E}_2 = \hat{\mathbf{y}} a_2 \cos(kz - \omega t + \theta) \quad (17)$$

For $\theta = n\pi$, the resultant will also be a linearly polarized wave with its electric vector oscillating along a direction making a certain angle with the x -axis; this angle will depend on the relative values of a_1 and a_2 .

In order to find the state of polarization of the resultant field, we consider the time variation of the resultant electric field at an arbitrary plane perpendicular to the z -axis which we may, without any loss of generality, assume to be $z = 0$. If E_x and E_y represent the x - and y -components of the resultant field $\mathbf{E} (= \mathbf{E}_1 + \mathbf{E}_2)$, then

$$E_x = a_1 \cos \omega t \quad (18)$$

and

$$E_y = a_2 \cos (\omega t - \theta) \quad (19)$$

where we have used Eqs (16) and (17) with $z = 0$. For $\theta = n\pi$, the above equations simplify to

$$E_x = a_1 \cos \omega t$$

$$\text{and} \quad E_y = (-1)^n a_2 \cos \omega t \quad (20)$$

from which one obtains

$$\frac{E_y}{E_x} = \pm \frac{a_2}{a_1} \quad (\text{independent of } t) \quad (21)$$

where the upper and lower signs correspond to n even and n odd respectively. In the E_x - E_y plane, Eq. (21) represents a straight line; the angle (ϕ) that this line makes with the E_x axis depends on the ratio a_2/a_1 . In fact

$$\phi = \tan^{-1} \left(\pm \frac{a_2}{a_1} \right) \quad (22)$$

The condition $\theta = n\pi$ implies that the two vibrations are either in phase ($n = 0, 2, 4, \dots$) or out of phase ($n = 1, 3, 5, \dots$). Thus, the superposition of two linearly polarized electromagnetic waves with their electric fields at right angles to each other and oscillating in phase, is again a linearly polarized wave with its electric vector, in general, oscillating in a direction which is different from the fields of either of the two waves. Figure 19.14 is a plot of the resultant field corresponding to Eq. (20) for various values of a_2/a_1 . The tip of the electric vector oscillates (with angular frequency ω) along the thick lines shown in the figure. The equation of the straight line is given by Eq. (21).

For $\theta \neq n\pi$ ($n = 0, 1, 2, \dots$), the resultant electric vector does not, in general, oscillate along a straight line. We first consider the simple case corresponding to $\theta = \pi/2$ with $a_1 = a_2$. Thus,

$$E_x = a_1 \cos \omega t \quad (23)$$

$$E_y = a_1 \sin \omega t \quad (24)$$

If we plot the time variation of the resultant electric vector whose x - and y -components are given by Eqs (23) and (24), one would find that the tip of the electric vector rotates

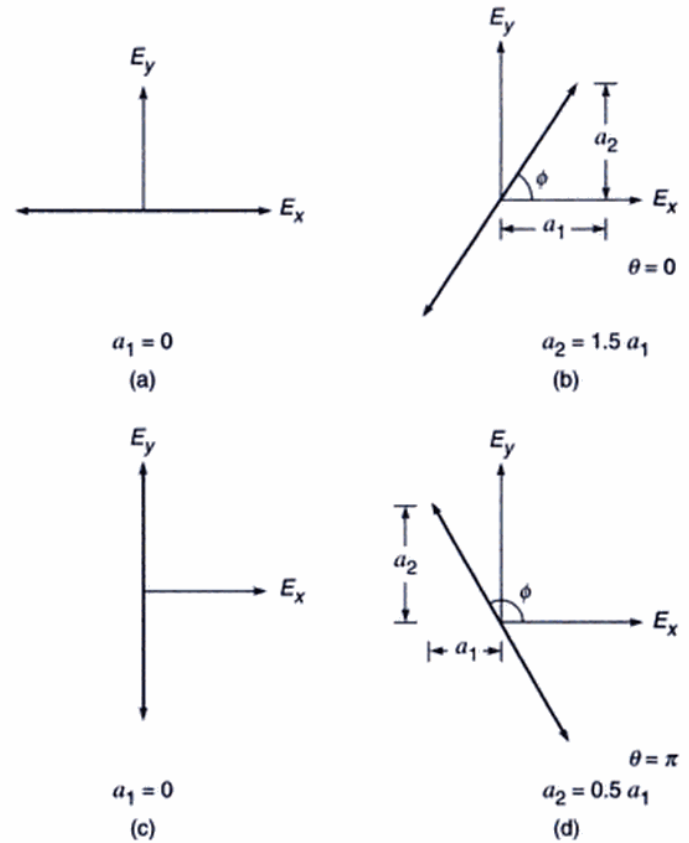


Fig. 19.14 The superposition of two linearly polarized waves with their electric fields oscillating in phase along the x -axis and the y -axis. The resultant is again a linearly polarized wave with its electric vector oscillating in a direction making an angle ϕ with the x -axis.

on the circumference of a circle (of radius a_1) in the anti-clockwise direction [see Fig. 19.15(c)] and the propagation is in the $+z$ -direction which is coming out of the page. Such a wave is known as a right circularly polarized wave (usually abbreviated as a RCP wave).^{*} That the tip of the resultant electric vector should lie on the circumference of a circle is also obvious from the fact that

$$E_x^2 + E_y^2 = a_1^2 \quad (\text{independent of } t) \quad (25)$$

For $\theta = 3\pi/2$, one would have had

$$E_x = a_1 \cos \omega t \quad (26)$$

$$E_y = -a_1 \sin \omega t \quad (27)$$

which would also represent a circularly polarized wave; however, the electric vector will rotate in the clockwise

^{*}Our convention for labelling left and right circularly polarized light is consistent with the one used by Ref. 4 but in some books the opposite convention is used.

In Sec. 19.12 we will show that whereas the velocity of the ordinary ray is the same in all directions, the velocity of the extraordinary ray is different in different directions; a substance (like calcite, quartz), which exhibits different properties in different directions, is called an anisotropic substance. Along a particular direction (fixed in the crystal), the two velocities are equal; this direction is known as the optic axis of the crystal. In a crystal like calcite, the two rays have the same speed only along one direction (which is the optic axis); such crystals are known as uniaxial crystals.* The velocities of the ordinary and the extra-ordinary rays are given by the following equations [see Eq. (121)]:

$$v_{ro} = \frac{c}{n_o} \quad (\text{ordinary ray}) \quad (34)$$

$$\frac{1}{v_{re}^2} = \frac{\sin^2 \theta}{(c/n_e)^2} + \frac{\cos^2 \theta}{(c/n_o)^2} \quad (\text{extra-ordinary ray}) \quad (35)$$

where n_o and n_e are constants of the crystal and θ is the angle that the ray makes with the optic axis; we have assumed the optic axis to be parallel to the z -axis. Thus, c/n_o and c/n_e are the velocities of the extraordinary ray when it propagates parallel and perpendicular to the optic axis. Now, the equation of an ellipse (in the z - x plane) is given by

$$\frac{z^2}{a^2} + \frac{x^2}{b^2} = 1 \quad (36)$$

If (ρ, θ) represents the polar coordinates then $z = \rho \cos \theta$ and $x = \rho \sin \theta$ and the equation of the ellipse can be written in the form

$$\frac{1}{\rho^2} = \frac{\cos^2 \theta}{a^2} + \frac{\sin^2 \theta}{b^2} \quad (37)$$

In three dimensions, Eq. 37 will represent an ellipsoid of revolution with the optic axis as the axis of revolution. Thus if we plot v_{re} as a function of θ , we will obtain an ellipsoid of revolution; on the other hand, since v_{ro} is independent of θ , if we plot v_{ro} (as a function of θ), we will obtain a sphere. Along the optic axis, $\theta = 0$ and

$$v_{ro} = v_{re} = \frac{c}{n_o}$$

We next consider the value of v_{re} perpendicular to the optic axis (i.e., for $\theta = \pi/2$). For a negative crystal $n_e < n_o$ and

$$v_{re} \left(\theta = \frac{\pi}{2} \right) = \frac{c}{n_e} > v_{ro} \quad (38)$$

Thus the minor axis will be along the optic axis and the ellipsoid of revolution will lie outside the sphere [see

Fig. 19.17(a)]. On the other hand, for a positive crystal $n_e > n_o$ and

$$v_{re} \left(\theta = \frac{\pi}{2} \right) = \frac{c}{n_e} < v_{ro} \quad (39)$$

The major axis will now be along the optic axis and the ellipsoid of revolution will lie inside the sphere [see Fig. 19.17(b)]. The ellipsoid of revolution and the sphere are known as the *ray velocity surfaces*.

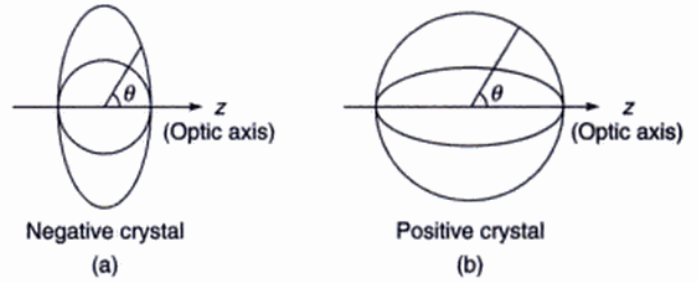


Fig. 19.17 (a) In a negative crystal, the ellipsoid of revolution (which corresponds to the extraordinary ray) lies outside the sphere; the sphere corresponds to the ordinary ray. (b) In a positive crystal, the ellipsoid of revolution (which corresponds to the extraordinary ray) lies inside the sphere.

We next consider an unpolarized plane wave incident on a calcite crystal. The plane wave will split up into two plane waves. One is referred to as the ordinary wave (usually abbreviated as the *o*-wave) and the other is referred to as the extra-ordinary wave (usually abbreviated as the *e*-wave). For both waves, the space and time dependence of the vectors \mathbf{E} , \mathbf{D} , \mathbf{B} and \mathbf{H} can be assumed to be of the form

$$e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}$$

where \mathbf{k} denotes the propagation vector and represents the direction normal to the phase fronts. In general, the \mathbf{k} vector for the *o*- and *e*-waves will be different. In Sec. 19.12 we will show that

1. Both ordinary and extra-ordinary waves are linearly polarized.

2. $\mathbf{D} \cdot \mathbf{k} = 0$ for both *o*- and *e*-waves (40)

Thus \mathbf{D} is always at right angles to \mathbf{k} and for this reason the direction of \mathbf{D} is chosen as the direction of 'vibrations'.

3. If we assume the z -axis to be parallel to the optic axis then

$$\mathbf{D} \cdot \hat{\mathbf{z}} = 0 \quad (\text{and } \mathbf{D} \cdot \mathbf{k} = 0) \quad \text{for the } o\text{-wave} \quad (41)$$

*In general, there may be two directions along which the two rays have the same speed; such crystals are known as biaxial crystals. The analysis of biaxial crystals is quite difficult; interested readers may look up Refs 5 and 6.

incident wavefront. If the time taken for the disturbance to reach the point F from D is t , then with B as centre we draw a sphere of radius $(c/n_o)t$ and an ellipsoid of revolution of semi-minor and semi-major axes $(c/n_o)t$ and $(c/n_e)t$ respectively; the semi-minor axis is along the optic axis. From the point F we draw tangent planes FO and FE to the sphere and the ellipsoid of revolution respectively. These planes would represent the refracted wavefronts corresponding to the ordinary and the extra-ordinary rays respectively. If the points of contact are O and E , then the ordinary and extra-ordinary refracted rays will propagate along BO and BE respectively; this can also be shown using Fermat's principle (see Sec. 2.5). The directions of vibration of these rays are shown by dots and small lines respectively and are obtained by using the general rules discussed earlier. The shape of the refracted wavefronts corresponding to the particular case of $\alpha = 0$ and $\alpha = \pi/2$ can be obtained very easily.

Figure 19.20(b) corresponds to the case when the optic axis is normal to the plane of incidence. The sections of both the wavefronts will be circles; consequently, the extra-ordinary ray will also satisfy Snell's law and we will have

$$\frac{\sin i}{\sin r} = n_e \quad (\text{for the } e\text{-ray when the optic-axis is normal to the plane of incidence}) \quad (44)$$

Of course, for the ordinary ray we will always have

$$\frac{\sin i}{\sin r} = n_o \quad (45)$$

19.6 INTERFERENCE OF POLARIZED LIGHT: QUARTER WAVE PLATES AND HALF WAVE PLATES

In the previous section we had considered how a plane wave (incident on a doubly refracting crystal) splits up into two waves each characterised by a certain state of polarization. The direction of vibration associated with the ordinary and extra-ordinary waves is obtained by using the recipe given by Eqs (41) and (42). In this section, we will consider the normal incidence of a plane-polarized beam on a calcite crystal whose optic axis is parallel to the surface of the crystal as shown in Fig. 19.21. We will study the state of polarization of the beam emerging from the crystal. We will assume the z -axis to be along the optic axis. Now, as discussed in the previous section, if the incident beam is y -polarized the beam will propagate as an ordinary wave and the extra-ordinary wave will be absent. Similarly, if the incident beam is z -polarized the beam will propagate as an extra-ordinary wave and the ordinary wave will be absent.

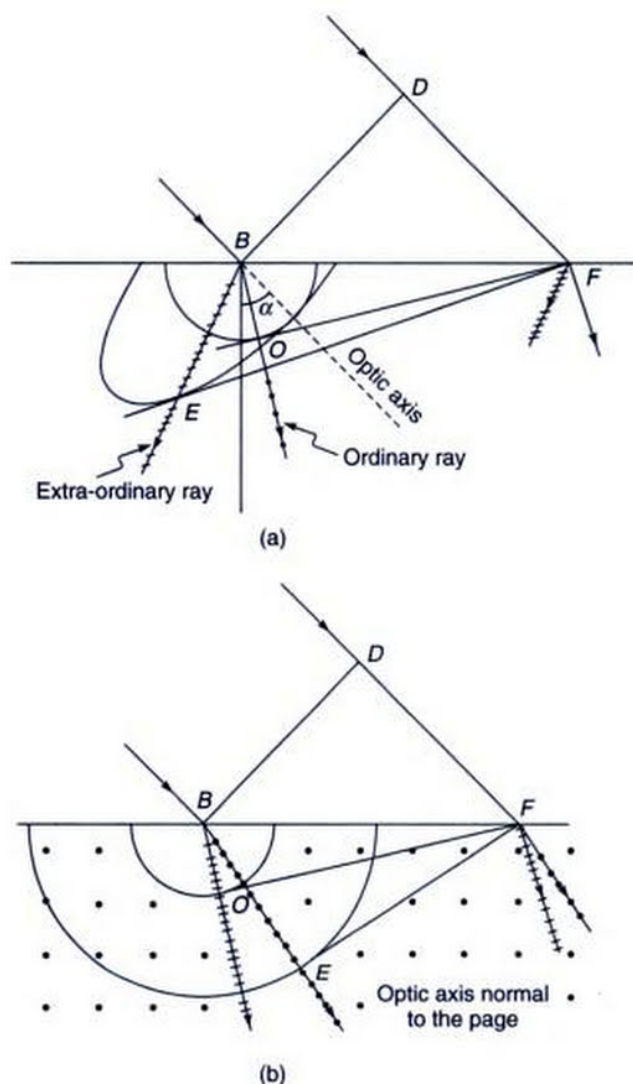


Fig. 19.20 Refraction of a plane wave incident obliquely on a negative uniaxial crystal. In (a), the direction of the optic axis is along the dashed line. In (b), the optic axis is perpendicular to the plane of the paper.

For any other state of polarization of the incident beam, both the extra-ordinary and the ordinary components will be present. For a negative crystal like calcite $n_e < n_o$ and the e -wave will travel faster than the o -wave; this is shown by putting s (slow) and f (fast) inside the parenthesis in Fig. 19.21.

Let the electric vector (of amplitude E_0) associated with the incident polarized beam make an angle ϕ with the z -axis; in Fig. 19.21 ϕ has been shown to be equal to 45° —but for the time being we will keep our analysis general and assume ϕ to be an arbitrary angle. Such a beam can be assumed to be a superposition of two linearly polarized beams (vibrating in phase), polarized along the y - and z -directions with

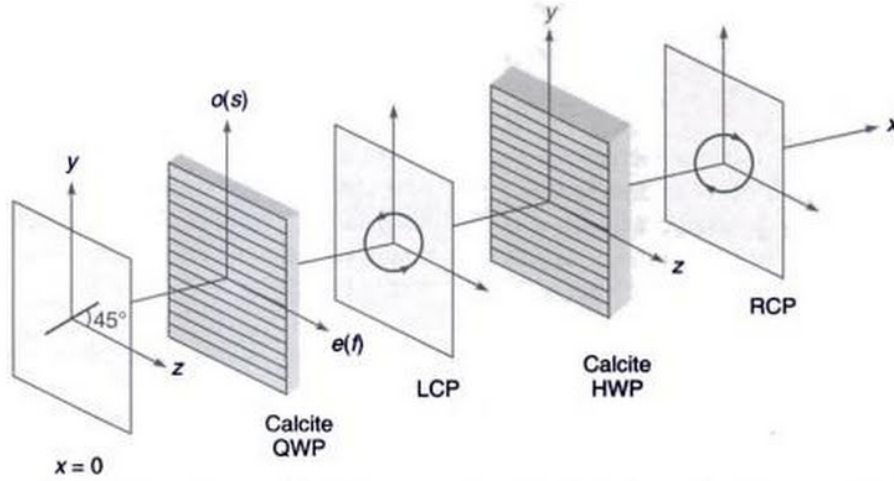


Fig. 19.21 A linearly polarized beam making an angle 45° with the z -axis gets converted to a LCP after propagating through a calcite QWP; further, an LCP gets converted to a RCP after propagating through a calcite HWP. The optic axis in the QWP and HWP is along the z -direction as show by lines parallel to the z -axis.

amplitudes $E_0 \sin \phi$ and $E_0 \cos \phi$ respectively. The z -component (whose amplitude is $E_0 \cos \phi$) passes through as an extra-ordinary beam propagating with wave velocity c/n_e . The y -component (whose amplitude is $E_0 \sin \phi$) passes through as an ordinary beam propagating with wave velocity c/n_o . Since $n_e \neq n_o$, the two beams will propagate with different velocities and, as such, when they come out of the crystal, they will not be in phase. Consequently, the emergent beam (which will be a superposition of these two beams) will be, in general, elliptically polarized.

Let the plane $x = 0$ represent the surface of the crystal on which the beam is incident. The y - and z -components of the incident beam can be written in the form

$$\left. \begin{aligned} E_y &= E_0 \sin \phi \cos(kx - \omega t) \\ E_z &= E_0 \cos \phi \cos(kx - \omega t) \end{aligned} \right\} \quad (46)$$

where $k (= \omega/c)$ represents the free-space wave number. Thus, at $x = 0$, we will have

$$\left. \begin{aligned} E_y(x=0) &= E_0 \sin \phi \cos \omega t \\ E_z(x=0) &= E_0 \cos \phi \cos \omega t \end{aligned} \right\}$$

Inside the crystal, the two components will be given by

$$\begin{aligned} E_y &= E_0 \sin \phi \cos(n_o kx - \omega t) \quad (\text{ordinary wave}) \\ E_z &= E_0 \cos \phi \cos(n_e kx - \omega t) \quad (\text{extra-ordinary wave}) \end{aligned}$$

If the thickness of the crystal is d , then at the emerging surface, we will have

$$\left. \begin{aligned} E_y &= E_0 \sin \phi \cos(\omega t - \theta_o) \\ E_z &= E_0 \cos \phi \cos(\omega t - \theta_e) \end{aligned} \right\}$$

where $\theta_o = n_o k d$ and $\theta_e = n_e k d$. By appropriately choosing the instant $t = 0$, the components may be rewritten as

$$\left. \begin{aligned} E_y &= E_0 \sin \phi \cos(\omega t - \theta) \\ E_z &= E_0 \cos \phi \cos \omega t \end{aligned} \right\} \quad (47)$$

where

$$\theta = \theta_o - \theta_e = k d (n_o - n_e) = \frac{\omega}{c} (n_o - n_e) d \quad (48)$$

represents the phase difference between the ordinary and the extra-ordinary beams. Clearly, if the thickness of the crystal is such that $\theta = 2\pi, 4\pi, \dots$, the emergent beam will have the same state of polarization as the incident beam. Now, if the thickness d of the crystal is such that $\theta = \pi/2$, the crystal is said to be a quarter wave plate (usually abbreviated as QWP)—a phase difference of $\pi/2$ implies a path difference of a quarter of a wavelength. On the other hand, if the thickness of the crystal is such that $\theta = \pi$, the crystal is said to be a half wave plate (usually abbreviated as HWP).

As an example, let us consider the case when $\phi = \pi/4$ and $\theta = \pi/2$, i.e., the y - and z -components of the incident wave have equal amplitudes and the crystal introduces a phase difference of $\pi/2$ (see Fig. 19.21). Thus, for the emergent beam we have

$$E_y = \frac{E_0}{\sqrt{2}} \sin \omega t; \quad E_z = \frac{E_0}{\sqrt{2}} \cos \omega t \quad (49)$$

which represents a circularly polarized wave because

$$E_y^2 + E_z^2 = \frac{E_0^2}{2}$$

In order to determine the direction of rotation of the electric vector, we note that at $t = 0$,

$$E_y = 0, \quad E_z = \frac{E_0}{\sqrt{2}}$$

and at $t = \Delta t$,

$$E_y = \frac{E_0}{\sqrt{2}} \omega \Delta t, \quad E_z = \frac{E_0}{\sqrt{2}}$$

The above equations show that as time increases, the electric vector rotates in the anti-clockwise direction and hence the beam is left circularly polarized as shown in Fig. 19.21. In order to introduce a phase difference of $\pi/2$, the thickness of the crystal should have a value given by the following equation:

$$d = \frac{c}{\omega(n_o - n_e)} \frac{\pi}{2} = \frac{1}{4} \frac{\lambda_0}{(n_o - n_e)} \quad (50)$$

where λ_0 is the free-space wavelength. For calcite,

$$n_o = 1.65836, \quad n_e = 1.48641$$

which correspond to $\lambda_0 = 5893 \text{ \AA}$ and at 18°C . Substituting these values, we obtain

$$d = \frac{5893 \times 10^{-8}}{4 \times 0.17195} \text{ cm} = 0.000857 \text{ mm}$$

Thus a calcite QWP (at $\lambda_0 = 5893 \text{ \AA}$) will have a thickness of 0.000857 mm and will have its optic axis parallel to the surface; such a QWP will introduce a phase difference of $\pi/2$ between the ordinary and the extra-ordinary components at $\lambda_0 = 5893 \text{ \AA}$. It should be pointed out that if the thickness is an odd multiple of the above quantity, i.e., if

$$d = (2m + 1) \frac{\lambda_0}{4(n_o - n_e)}; \quad m = 0, 1, 2, \dots \quad (51)$$

then in the example considered above (i.e., when $\phi = \pi/4$), it can easily be shown that the emergent beam will be left circularly polarized for $m = 0, 2, 4, \dots$ and right circularly polarized for $m = 1, 3, 4, \dots$. It may be mentioned that the y-polarized o-wave in calcite has a smaller wave velocity

($= c/n_o$) and hence it is referred to as a slow wave hence shown as $o(s)$ in Figs. 19.21 and 19.22; similarly, the extraordinary wave is the fast wave (in calcite) hence shown as $e(f)$.

We next consider the case when the linearly polarized beam (with $\phi = \pi/4$) is incident on a HWP so that $\theta = \pi$, i.e., the y- and z-components of the incident wave have equal amplitudes and the crystal introduces a phase difference of π (see Fig. 19.22). Thus, for the emergent beam we have

$$E_y = -\frac{E_0}{\sqrt{2}} \cos \omega t; \quad E_z = \frac{E_0}{\sqrt{2}} \cos \omega t$$

which represents a linearly polarized wave with the direction of polarization making an angle of 135° with the z-axis (see Fig. 19.22). If we now pass this beam through a calcite QWP, the emergent beam will be right circularly polarized as shown in Fig. 19.22. On the other hand, if a left circularly polarized is incident normally on a calcite HWP, the emergent beam will be right circularly polarized as shown in Fig. 19.21.

Thus, for a HWP the thickness (for a negative crystal) would be given by

$$d = (2m + 1) \frac{\lambda_0}{2(n_o - n_e)}$$

We may mention that if the crystal thickness is such that if $\theta \neq \pi/2, \pi, 3\pi/2, 2\pi, \dots$ the emergent beam will be elliptically polarized; similar to that shown in Fig. 19.15 (of course, there the propagation was along the z-axis and here it is along the x-axis).

For a positive crystal (like quartz), $n_e > n_o$ and Eq. (47) should be written in the form

$$\left. \begin{aligned} E_y &= E_0 \sin \phi \cos(\omega t + \theta') \\ E_z &= E_0 \cos \phi \cos \omega t \end{aligned} \right\} \quad (52)$$

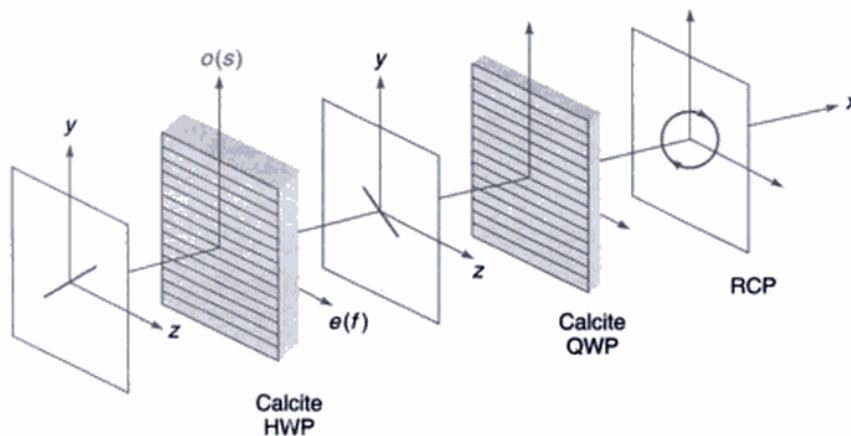


Fig. 19.22 If the linearly polarized beam making an angle 45° with the z-axis is incident on a HWP, the plane of polarization gets rotated by 90° ; this beam gets converted to a RCP after propagating through a calcite QWP. The optic axis in the HWP and QWP is along the z-direction as shown by lines parallel to the z-axis.

where

$$\theta' = \frac{\omega}{c} d (n_e - n_o).$$

For a quarter wave plate,

$$d = (2m + 1) \frac{\lambda_0}{4(n_e - n_o)}; \quad m = 0, 1, 2, \dots$$

Thus, if in Fig. 19.21, the calcite QWP is replaced by a quartz QWP, the emergent beam will be right circularly polarized.

Example 19.1 A left circularly polarized beam ($\lambda_0 = 5893 \text{ \AA}$) is incident normally on a calcite crystal (with its optic axis cut parallel to the surface) of thickness 0.005141 mm . What will be the state of polarization of the emergent beam?

Solution: The electric field for the incident beam at $x = 0$ would be

$$E_y = \frac{E_0}{\sqrt{2}} \sin \omega t; \quad E_z = \frac{E_0}{\sqrt{2}} \cos \omega t \quad (53)$$

Now

$$\begin{aligned} \theta &= \frac{(n_o - n_e) d \times 2\pi}{\lambda_0} \\ &= \frac{0.17195 \times 0.005141 \times 2\pi}{5893 \times 10^{-7}} \approx 3\pi \end{aligned}$$

Thus the emergent wave will be [cf. Eq. (47)]

$$\begin{aligned} E_y &= \frac{E_0}{\sqrt{2}} \sin(\omega t - 3\pi) \\ &= -\frac{E_0}{\sqrt{2}} \sin \omega t; \quad E_z = \frac{E_0}{\sqrt{2}} \cos \omega t \end{aligned}$$

which represents a right circularly polarized beam.

Example 19.2 A left circularly polarized beam ($\lambda_0 = 5893 \text{ \AA}$) is incident on a quartz crystal (with its optic axis cut parallel to the surface) of thickness 0.025 mm . Determine the state of polarization of the emergent beam. Assume n_o and n_e to be 1.54425 and 1.55336 respectively.

Solution: As in the previous example, the electric field for the incident beam at $x = 0$ would be given by Eq. (53). Further,

$$\theta' = (n_e - n_o) \frac{2\pi}{\lambda_0} d = 2\pi \frac{0.00911 \times 0.025}{5893 \times 10^{-7}} = 0.77\pi$$

Thus the emergent beam will be

$$E_y = \frac{E_0}{\sqrt{2}} \cos(\omega t + 0.77\pi); \quad E_z = \frac{E_0}{\sqrt{2}} \cos(\omega t)$$

which will represent a right elliptically polarized light.

19.7 ANALYSIS OF POLARIZED LIGHT

In the previous sections we have seen that a plane wave can be characterized by different states of polarizations, which may be any one of the following:

- linearly polarized,
- circularly polarized,
- elliptically polarized,
- unpolarized,
- mixture of linearly polarized and unpolarized,
- mixture of circularly polarized and unpolarized, or
- mixture of elliptically polarized and unpolarized light.

To the naked eye, all the states of polarizations will appear to be the same. In this section, we will discuss the procedure for determining the state of polarization of a light beam.

If we introduce a polaroid in the path of the beam and rotate it about the direction of propagation, then either of the following three possibilities can occur:

- If there is complete extinction at two positions of the polarizer, then the beam is linearly polarized.
- If there is no variation of intensity, then the beam is either unpolarized or circularly polarized or a mixture of unpolarized and circularly polarized light. We now put a quarter wave plate on the path of the beam followed by the rotating polaroid. If there is no variation of intensity then the incident beam is unpolarized. If there is complete extinction at two positions, then the beam is circularly polarized (this is due to the fact that a quarter wave plate will transform a circularly polarized light into a linearly polarized light). If there is a variation of intensity (without complete extinction) then the beam is a mixture of unpolarized and circularly polarized light.
- If there is a variation of intensity (without complete extinction) then the beam is either elliptically polarized or a mixture of linearly polarized and unpolarized or a mixture of elliptically polarized and unpolarized light. We now put a quarter wave plate in front of the polaroid with its optic axis parallel to the pass-axis of the polaroid at the position of maximum intensity. The elliptically polarized light will transform to a linearly polarized light. Thus, if one obtains two positions of the polaroid where complete extinction occurs, then the original beam is elliptically polarized. If complete extinction does not occur, and the position of maximum intensity occurs at the same orientation as before, the beam is a mixture of unpolarized and linearly polarized light. Finally, if the position of maximum intensity occurs at a different orientation of the polaroid, the beam is a mixture of elliptically polarized and unpolarized light.

19.8 OPTICAL ACTIVITY

When a linearly polarized light beam propagates through an 'optically active' medium like sugar solution then—as the beam propagates—its plane of polarization rotates. This rotation is directly proportional to the distance traversed by the beam and also to the concentration of sugar in the solution. Indeed, by measuring the angle by which the plane of polarization is rotated, one can accurately determine the concentration of sugar in the solution.

The rotation of the plane of polarization is due to the fact that the 'modes' of the optically active substance are left circularly polarized (LCP) and right circularly polarized (RCP) which propagate with slightly different velocities. By 'modes' we imply that if an LCP light beam is incident on the substance then it will propagate as an LCP beam; similarly, an RCP light beam will propagate as an RCP beam but with a slightly different velocity. On the other hand, if a linearly polarized light beam is incident, then we must express the linear polarization as a superposition of an RCP and an LCP beam and then consider the independent propagation of the two beams. We illustrate through an example.

We consider an RCP beam propagating in the $+z$ direction

$$\left. \begin{aligned} E_x^r &= E_0 \cos(k_r z - \omega t) \\ E_y^r &= -E_0 \sin(k_r z - \omega t) \end{aligned} \right\} \quad (54)$$

where $k_r = \frac{\omega}{c} n_r$ and the superscript (and the subscript) r signify that we are considering an RCP beam. Similarly, an LCP beam (of the same amplitude) propagating in the $+z$ direction can be described by the following equations:

$$\left. \begin{aligned} E_x^l &= E_0 \cos(k_l z - \omega t) \\ E_y^l &= E_0 \sin(k_l z - \omega t) \end{aligned} \right\} \quad (55)$$

where $k_l = \frac{\omega}{c} n_l$; n_r and n_l are the refractive indices corresponding to the RCP and LCP beams respectively. If we assume the simultaneous propagation of the two beams then the x and y components of the resultant fields would be given by the following equations:

$$E_x = E_0 [\cos(k_r z - \omega t) + \cos(k_l z - \omega t)]$$

or

$$E_x = 2E_0 \cos \left[\frac{1}{2} (k_l - k_r) z \right] \cos[\omega t - \theta(z)]$$

Similarly

$$E_y = 2E_0 \sin \left[\frac{1}{2} (k_l - k_r) z \right] \cos[\omega t - \theta(z)]$$

*When a wave propagates along the optic axis of a quartz crystal it is strictly speaking, not like calcite. The modes are *not* linearly polarized; they are RCP and LCP propagating with slightly different velocities.

where

$$\theta(z) = \frac{1}{2} (k_l + k_r) z$$

Thus the resultant wave is *always* linearly polarized with the plane of polarization rotating with z . If the direction of the oscillating electric vector makes an angle ϕ with the x -axis then (see Fig. 19.23):

$$\phi(z) = \frac{1}{2} (k_l - k_r) z$$

or

$$\phi(z) = \frac{\pi}{\lambda_0} (n_l - n_r) z = \frac{\omega}{2c} (n_l - n_r) z \quad (56)$$

where λ_0 is the free space wavelength. Now, if

$n_l > n_r \Leftrightarrow$ the optically active substance is said to be right-handed or dextro-rotatory

$n_r > n_l \Leftrightarrow$ the optically active substance is said to be left-handed or laevo-rotatory

For example, for turpentine,

$$\phi = +37^\circ \text{ for } z = 10 \text{ cm}$$

As mentioned earlier, we observe optical activity even in a sugar solution, and this is due to the helical structure of sugar molecules. The method of determining the concentration of sugar solutions by measuring the rotation of the plane of polarization is a widely used method in industry. It may be noted that if $n_l = n_r$ (as is indeed the case in an isotropic substance) then $\phi(z) = 0$ and a linearly polarized beam remains linearly polarized along the same direction. Optical activity is also exhibited in crystals. For example, for a linearly polarized light propagating along the optic axis of a quartz crystal*, the plane of polarization gets rotated. Indeed

$$|n_l - n_r| \approx 7 \times 10^{-5}$$

$$\Rightarrow \phi \approx \frac{7}{60} \pi = 21^\circ$$

$$\text{for } z = 0.1 \text{ cm at } \lambda_0 = 6000 \text{ \AA}$$

19.9 CHANGE IN THE SOP (STATE OF POLARIZATION) OF A LIGHT BEAM PROPAGATING THROUGH AN ELLIPTIC CORE SINGLE MODE OPTICAL FIBER

A very interesting phenomenon is the propagation of polarized light through an elliptic core optical fiber. We will

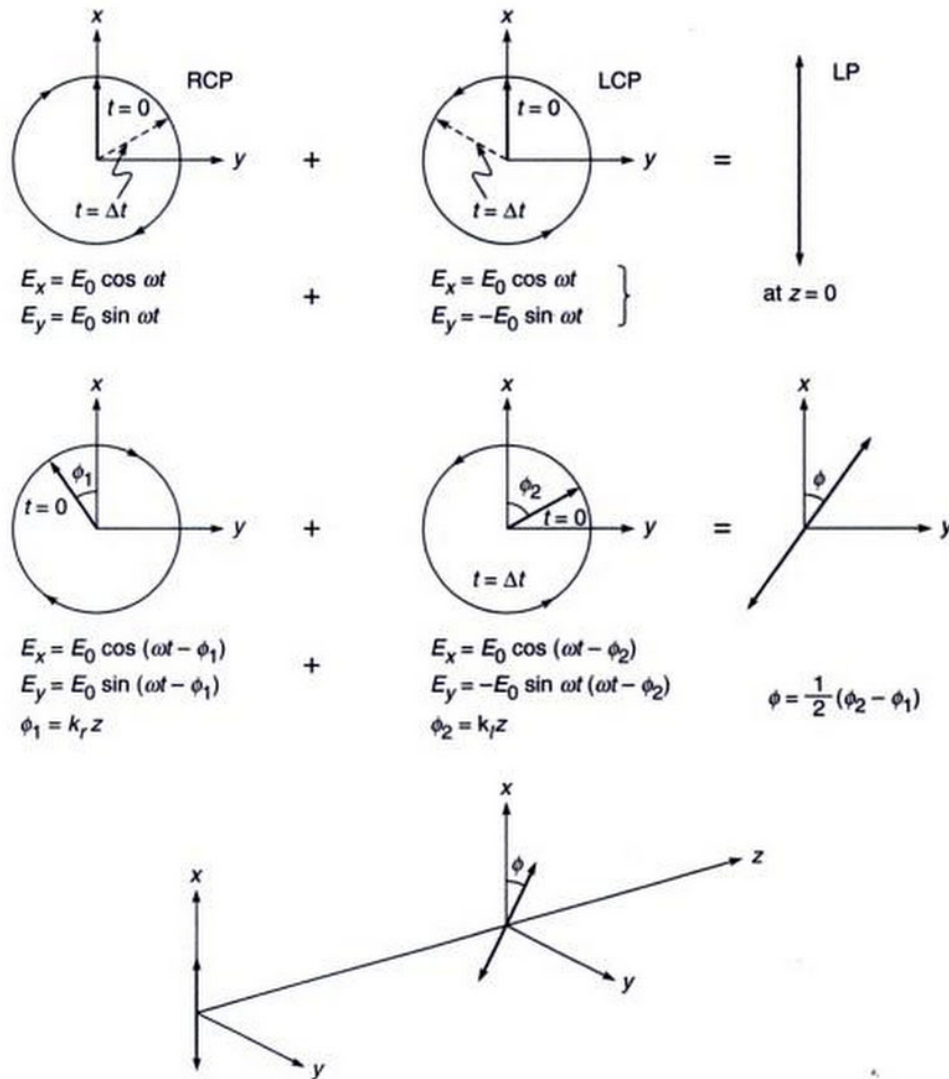


Fig. 19.23 The "clockwise" rotation of a linearly polarized wave as it propagates through a "right-handed" optically active medium.

have a brief discussion on optical fibers in Chapter 24; it will suffice here to say that in an ordinary optical fiber we have a cylindrical core (of circular cross-section) clad with a medium of slightly lower refractive index. The guidance of the light beam takes place through the phenomenon of total internal reflection (see Figs. 24.2 and 24.7). Because of the circular symmetry of the problem, the incident beam can have any state of polarization* which will be maintained as the beam propagates through the fiber. Now, if we have an elliptic core fiber [see Fig. 19.24(a)] then the 'modes' of the fiber are (approximately) x and y polarized; i.e., if an x -polarized beam is incident on the fiber, it will propagate without any change in the state of

polarization with a certain phase velocity ω/β_x . Similarly, a y -polarized beam will propagate as a y -polarized beam with velocity ω/β_y . Now, let a circularly polarized beam be incident on the input face of the fiber at $z=0$. Then we must resolve the incident beam into x and y polarized beams propagating with slightly different velocities. Thus

$$\mathbf{E}(x, y, z) = \psi(x, y) [\hat{x} \cos(\beta_x z - \omega t) + \hat{y} \sin(\beta_y z - \omega t)] \quad (57)$$

where $\psi(x, y)$ is the transverse field distribution of the fundamental mode which is assumed to be (approximately) the same for both x - and y -polarizations (see Sec. 24.9.1). It may be readily seen that if $\beta_x = \beta_y$, as is indeed true for

*We are considering here a single mode fiber so that no matter what the incident transverse field distribution is, it soon 'settles down' to the transverse field distribution of the fundamental mode which propagates with the velocity ω/β_0 . This velocity is independent of the SOP of the incident beam.

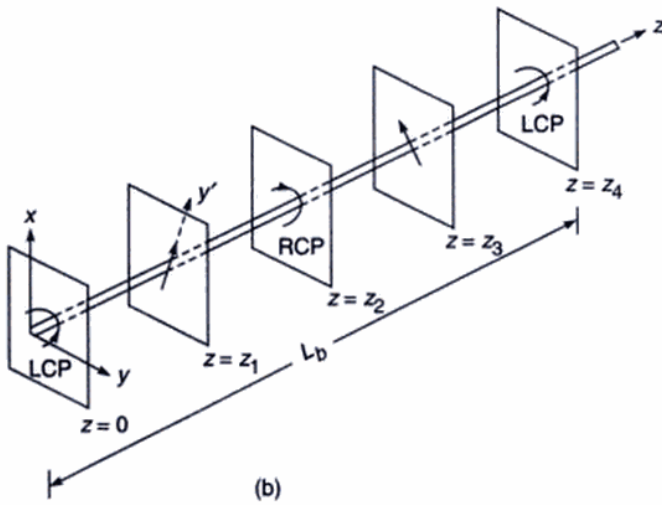
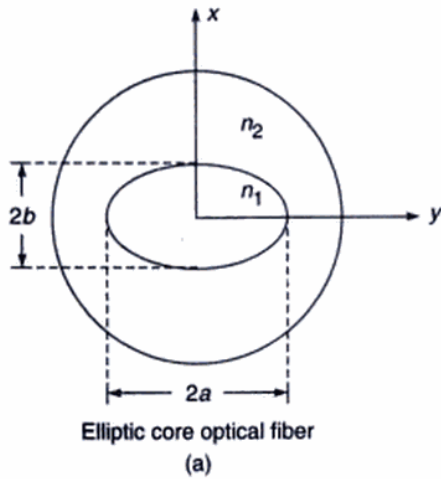


Fig. 19.24 (a) The transverse cross-section of an elliptic core fiber; the "modes" are (approximately) x -polarized and y -polarized. (b) Propagation of a left-circularly polarized beam incident on an elliptic core fiber. If we view along the y' -axis then dark spots will be observed at $z = z_1, 5z_1, 9z_1, \dots$

circular core fibers, the beam will remain circularly polarized for all values of z . Now, at $z = 0$,

$$\begin{aligned} E_x &= \psi(x, y) \cos \omega t \\ E_y &= -\psi(x, y) \sin \omega t \end{aligned} \quad (58)$$

which represents a left-circularly polarized wave [see Fig. 19.24(b)]. For

$$z = z_1 = \frac{\pi}{2(\beta_y - \beta_x)} \quad (59)$$

i.e., for $\beta_y z_1 = \beta_x z_1 + \pi/2$,

$$\begin{aligned} E_x &= \psi(x, y) \cos (\phi_1 - \omega t) \\ &= +\psi(x, y) \cos (\omega t - \phi_1) \end{aligned}$$

$$\begin{aligned} E_y &= \psi(x, y) \sin \left(\phi_1 + \frac{\pi}{2} - \omega t \right) \\ &= +\psi(x, y) \cos (\omega t - \phi_1) \end{aligned}$$

where

$$\phi_1 = \beta_x z_1$$

which represents a linearly polarized wave [see Fig. 19.24(b)]; we assume the direction of the \mathbf{E} -vector to be along the y' axis. Similarly, at

$$z = z_2 = \frac{\pi}{(\beta_y - \beta_x)} = 2z_1$$

$$\left. \begin{aligned} E_x &= \psi(x, y) \cos (\phi_2 - \omega t) \\ &= \psi(x, y) \cos (\omega t - \phi_2) \\ E_y &= \psi(x, y) \sin (\phi_2 + \pi - \omega t) \\ &= \psi(x, y) \sin (\omega t - \phi_2) \end{aligned} \right\} \quad (60)$$

where

$$\phi_2 = \beta_x z_2$$

and the wave will be right-circularly polarized [see Fig. 19.24(b)]. At

$$z = z_3 = \frac{3\pi}{2(\beta_x - \beta_y)} = 3z_1$$

we will have

$$\begin{aligned} E_x &= \psi(x, y) \cos (\phi_3 - \omega t) \\ &= \psi(x, y) \cos (\omega t - \phi_3) \\ E_y &= \psi(x, y) \sin \left(\phi_3 + \frac{3\pi}{2} - \omega t \right) \\ &= -\psi(x, y) \cos (\omega t - \phi_3) \end{aligned}$$

where

$$\phi_3 = \beta_x z_3$$

Thus the wave would again be linearly polarized but now the direction of the oscillating electric field will be at right angles to the field at $z = z_1$. In a similar manner, we can easily continue to determine the SOP of the propagating beam. Thus at $z = 5z_1, 9z_1, 13z_1, \dots$ the SOP will be the same as at $z = z_1$ and at $z = 7z_1, 11z_1, 15z_1, \dots$ the SOP will be the same as at $z = 3z_1$. Similarly at $z = 4z_1, 8z_1, 12z_1, \dots$ the beam will be LCP and at $z = 2z_1, 6z_1, 10z_1, \dots$ the beam will be RCP.

Now, let the fiber be rotated in such a way that the y' axis is along the vertical line (the x' and the z -axes are assumed to lie in the horizontal plane). Thus if we put our eyes vertically above the fiber and view vertically down then the regions $z = z_1, 5z_1, 9z_1, \dots$ will appear dark (see Fig. 19.25). This is because of the fact that in these regions the electric field is oscillating in the y' direction (which is the vertical

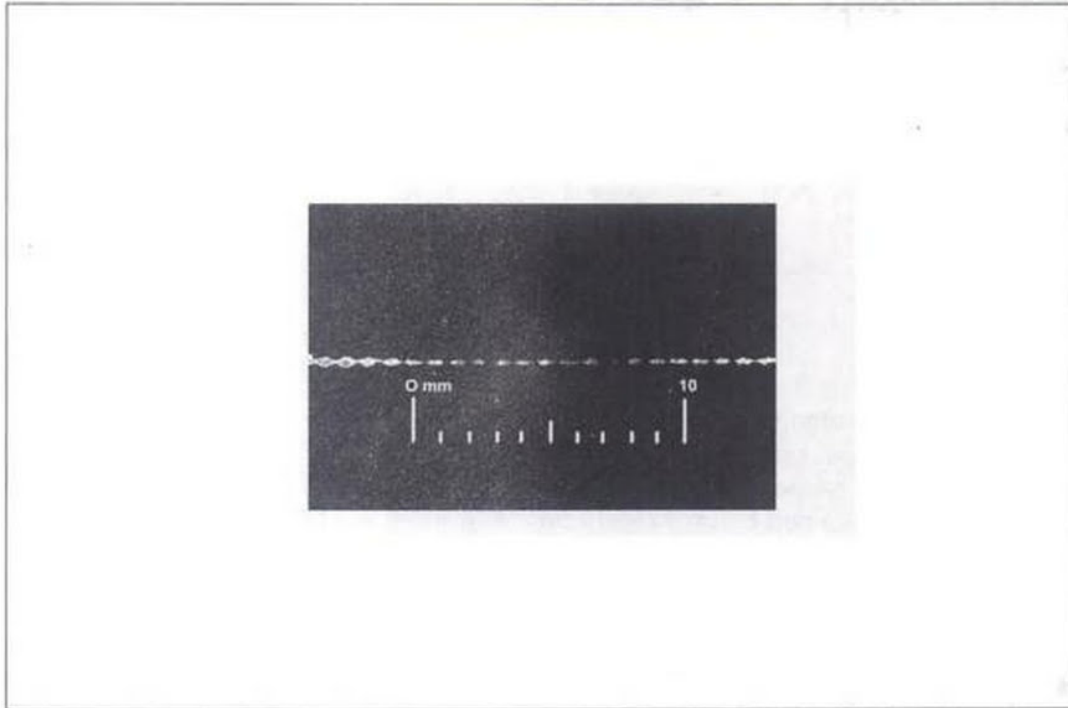


Fig. 19.25 A circularly polarized beam is incident normally on an elliptic core fiber—the photograph shows the intensity variation as seen from the top; adapted from a photograph by Andrew Corporation, Orland Park, IL USA; the author came across this photograph in Ref. 13.

direction) and we know that if the dipole oscillates along the y' direction, there is no radiation emitted in that particular direction (see Figs 19.12 and 20.4). Thus by measuring the distance between two consecutive black spots ($= 4z_1$) one can calculate z_1 and hence $\beta_y - \beta_x$. Furthermore, by moving the eyes to the horizontal plane, i.e., viewing along the x' axis we will see the regions $z = z_1, 5z_1, 9z_1, \dots$ appear bright and the regions $z = 3z_1, 7z_1, 11z_1, \dots$ appear dark. Thus the experiment not only allows one to understand the changing SOP of a beam propagating through a birefringent fiber, but also helps us understand the radiation pattern of an oscillating dipole.

As a numerical example, we consider an elliptical core fiber for which

$$2a = 2.14 \mu\text{m}, \quad 2b = 8.85 \mu\text{m} \\ n_1 = 1.535, \quad n_2 = 1.47$$

[see Fig. 19.24(a)]. For such a fiber operating at $\lambda_0 = 6328 \text{ \AA}$ ($k_0 \approx 9.929 \times 10^4 \text{ cm}^{-1}$),

$$\frac{\beta_x}{k_0} \approx 1.506845 \quad \text{and} \quad \frac{\beta_y}{k_0} \approx 1.507716$$

The quantity

$$L_b = \frac{2\pi}{\Delta\beta} = \frac{2\pi}{\beta_y - \beta_x} \approx 0.727 \text{ mm}$$

is known as the coupling length.

19.10 WOLLASTON PRISM

A Wollaston prism is used to produce two linearly polarized beams. It consists of two similar prisms (of say calcite) with the optic axis of the first prism parallel to the surface and the optic axis of the second prism parallel to the edge of the prism as shown in Fig. 19.26. Let us first consider the incidence of a z -polarized beam as shown in Fig. 19.26(a). The beam will propagate as an o -ray in the first prism (because the vibrations are perpendicular to the optic axis) and will see the refractive index n_o . When this beam enters the second prism it will become an e -ray and will see the refractive index n_e . For calcite $n_o > n_e$ and therefore the ray will bend away from the normal. Since the optic axis is normal to the plane of incidence, the refracted ray will obey Snell's laws [see Fig. 19.20(b)] and the angle of refraction will be given by

$$n_o \sin 20^\circ = n_e \sin r_1$$

where we have assumed the angle of the prism to be 20° (see Fig. 19.26). Assuming $n_o \approx 1.658$ and $n_e \approx 1.486$, we readily get

$$r_1 \approx 22.43^\circ$$

Thus the angle of incidence at the second surface will be $i_1 = 22.43^\circ - 20^\circ = 2.43^\circ$. The output angle θ_1 will be given by $n_e \sin 2.43^\circ = \sin \theta_1 \Rightarrow \theta_1 \approx 3.61^\circ$.

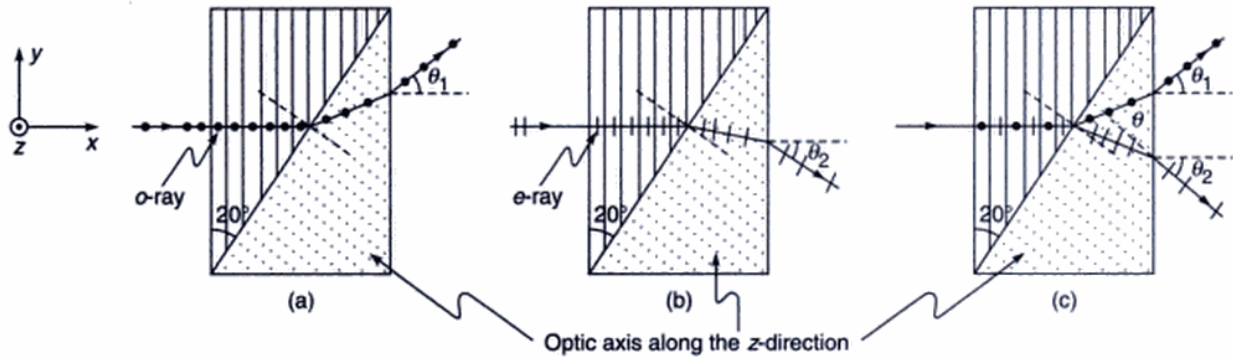


Fig. 19.26 A Wollaston prism. The optic axis of the first prism is along the y -axis and the optic axis of the second prism is along the z -axis. (a) If the incident beam is z -polarized, it will propagate as an o -wave in the first prism and an e -wave in the second prism. (b) If the incident beam is y -polarized, it will propagate as an e -wave in the first prism and an o -wave in the second prism. (c) For an unpolarized beam incident normally, there will be 2 linearly polarized beams propagating in different directions. The ray paths correspond to prisms being of calcite.

We next consider the incidence of a y -polarized beam as shown in Fig. 19.26(b). The beam will propagate as an e -ray in the first prism and as an o -ray in the second prism. The angle of refraction will now be given by

$$n_e \sin 20 = n_o \sin r_2 \Rightarrow r_2 = 17.85^\circ$$

Thus the angle of incidence at the second interface will be

$$i_2 = 20^\circ - 17.85^\circ = 2.15^\circ$$

The output angle θ_2 will be given by

$$n_o \sin 2.15^\circ = \sin \theta_2 \Rightarrow \theta_2 \approx 3.57^\circ$$

Thus, if an unpolarized beam is incident on the Wollaston prism, the angular separation between the two orthogonally polarized beams will be $\theta = \theta_1 + \theta_2 \approx 7.18^\circ$.

19.11 ROCHON PRISM

We next consider the Rochon prism which consists of two similar prisms of (say) calcite; the optic axis of the first prism is normal to the face of the prism while the optic axis of the second prism is parallel to the edge as shown in Fig. 19.27. Now, in the first prism both the beams will see the same refractive index n_o ; this follows from the fact that the ordinary and extra-ordinary waves travel with the same velocity ($= c/n_o$) along the optic axis of the crystal.

When the beam enters the second crystal, the ordinary ray (whose \mathbf{D} is normal to the optic axis) will see the same refractive index and go undeviated as shown in Fig. 19.27. On the other hand, the extra-ordinary ray (whose \mathbf{D} is along the optic axis) will see the refractive index n_e will bend away from the normal. The angle of refraction will be determined from the following equation

$$n_o \sin 18^\circ = n_e \sin r$$

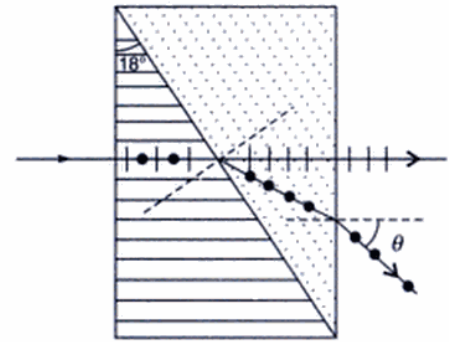


Fig. 19.27 Production of two orthogonally polarized beams by a Rochon prism.

$$\begin{aligned} \text{Thus } \sin r &= \frac{n_o}{n_e} \sin 18^\circ \\ &= \frac{1.658}{1.486} \times 0.309 \approx 0.345 \\ \Rightarrow r &= 20.2^\circ \end{aligned}$$

Therefore the angle of incidence at the second surface will be $20.2 - 18^\circ = 2.2^\circ$. The emerging angle will be given by

$$\begin{aligned} \sin \theta &= n_e \sin (2.2^\circ) \approx 0.057 \\ \Rightarrow \theta &\approx 3.3^\circ \end{aligned}$$

19.12 PLANE WAVE PROPAGATION IN ANISOTROPIC MEDIA

In this section, we will discuss the plane wave solutions of Maxwell's equations in an anisotropic medium and prove the various assumptions made in Sec. 19.5. The difference

between an isotropic and an anisotropic medium is in the relationship between the displacement vector \mathbf{D} and the electric vector \mathbf{E} ; the displacement vector \mathbf{D} is defined in Sec. 20.9. In an isotropic medium, \mathbf{D} is in the same direction as \mathbf{E} and one can write

$$\mathbf{D} = \epsilon \mathbf{E} \quad (61)$$

where ϵ is the dielectric permittivity of the medium. On the other hand, in an anisotropic medium \mathbf{D} is not, in general, in the direction of \mathbf{E} and the relation between \mathbf{D} and \mathbf{E} can be written in the form

$$\begin{cases} D_x = \epsilon_{xx} E_x + \epsilon_{xy} E_y + \epsilon_{xz} E_z \\ D_y = \epsilon_{yx} E_x + \epsilon_{yy} E_y + \epsilon_{yz} E_z \\ D_z = \epsilon_{zx} E_x + \epsilon_{zy} E_y + \epsilon_{zz} E_z \end{cases} \quad (62)$$

where $\epsilon_{xx}, \epsilon_{xy}, \dots$ are constants. One can show that*

$$\begin{aligned} \epsilon_{xy} &= \epsilon_{yx}; \quad \epsilon_{xz} = \epsilon_{zx} \\ \text{and} \quad \epsilon_{yz} &= \epsilon_{zy} \end{aligned} \quad (63)$$

Further, one can always choose a coordinate system (i.e., one can always choose appropriately the directions of x, y and z axes inside the crystal) such that

$$\begin{cases} D_x = \epsilon_x E_x \\ D_y = \epsilon_y E_y \\ \text{and} \quad D_z = \epsilon_z E_z \end{cases} \quad (64)$$

This coordinate system is known as the principal axis system and the quantities ϵ_x, ϵ_y and ϵ_z are known as the principle dielectric permittivities of the medium. If

$$\epsilon_x \neq \epsilon_y \neq \epsilon_z \quad (\text{biaxial}) \quad (65)$$

we have what is known as a biaxial medium and the quantities

$$n_x = \sqrt{\frac{\epsilon_x}{\epsilon_0}}, \quad n_y = \sqrt{\frac{\epsilon_y}{\epsilon_0}}, \quad n_z = \sqrt{\frac{\epsilon_z}{\epsilon_0}} \quad (66)$$

are said to be the principal refractive indices of the medium; in the above equation ϵ_0 represents the dielectric permittivity of free space ($= 8.8542 \times 10^{-12} \text{ C}^2/\text{N}\cdot\text{m}^2$). If

$$\epsilon_x = \epsilon_y \neq \epsilon_z \quad (\text{uniaxial}) \quad (67)$$

we have what is known as a uniaxial medium with the z -axis representing the optic axis of the medium. The quantities

$$n_o = \sqrt{\frac{\epsilon_x}{\epsilon_0}} = \sqrt{\frac{\epsilon_y}{\epsilon_0}}$$

and

$$n_e = n_z = \sqrt{\frac{\epsilon_z}{\epsilon_0}} \quad (68)$$

are known as ordinary and extra-ordinary refractive indices; typical values for some uniaxial crystals are given in Table 19.1. For a uniaxial medium, since $\epsilon_x = \epsilon_y$, the x and y directions can be arbitrarily chosen as long as they are perpendicular to the optic axis, i.e., any two mutually perpendicular axes (which are also perpendicular to the z -axis) can be taken as the principal axes of the medium.** On the other hand, if

$$\epsilon_x = \epsilon_y = \epsilon_z \quad (\text{isotropic}) \quad (69)$$

Table 19.1 Ordinary and extra-ordinary refractive indices for some uniaxial crystals (Table adapted from Ref. 6 and 7).

Name of the crystal	Wavelength	n_o	n_e
Calcite	4046 Å	1.68134	1.49694
	5890 Å	1.65835	1.48640
	7065 Å	1.65207	1.48359
Quartz	5890 Å	1.54424	1.55335
Lithium niobate	6000 Å	2.2967	2.2082
KDP	6328 Å	1.50737	1.46685
ADP	6328 Å	1.52166	1.47685

we have an isotropic medium, and can choose any three mutually perpendicular axes as the principal axis system. We will assume the anisotropic medium to be non-magnetic so that

$$\mathbf{B} = \mu_0 \mathbf{H}$$

where μ_0 is the free space magnetic permeability.

Let us consider the propagation of a plane electromagnetic wave; for such a wave the vectors $\mathbf{E}, \mathbf{H}, \mathbf{D}$ and \mathbf{B} would be proportional to $\exp[i(\mathbf{k} \cdot \mathbf{r} - \omega t)]$. Thus

$$\begin{cases} \mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} & \mathbf{H} = \mathbf{H}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \\ \mathbf{D} = \mathbf{D}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} & \mathbf{B} = \mathbf{B}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \end{cases} \quad (70)$$

where the vectors $\mathbf{E}_0, \mathbf{H}_0, \mathbf{D}_0$ and \mathbf{B}_0 are independent of space and time; \mathbf{k} represents the propagation vector of the

*See, e.g., Ref. 5.

**This follows from the fact that for a uniaxial medium

$$D_x = \epsilon_x E_x \text{ and } D_y = \epsilon_y E_y = \epsilon_x E_y$$

Now, if we rotate the x - y axes (about the z -axis) by an angle θ and call the rotated axes x' and y' , then

$$\begin{aligned} D_{x'} &= D_x \cos \theta + D_y \sin \theta = \epsilon_x [E_x \cos \theta + E_y \sin \theta] \\ &= \epsilon_x E_{x'} \end{aligned}$$

Similarly $D_{y'} = \epsilon_x E_{y'}$, implying that the $x' y'$ axes can also be chosen as principal axes.

wave and ω the angular frequency. The wave velocity v_w (also known as the phase velocity) and the wave refractive index n_w are defined through the following equation:

$$v_w = \frac{\omega}{k} = \frac{c}{n_w} \quad (71)$$

Thus

$$|\mathbf{k}| = k = \frac{\omega}{c} n_w \quad (72)$$

In the present section, it is our objective to determine the possible values of n_w when a plane wave propagates through an anisotropic dielectric. Now, in a dielectric medium

$$\text{div } \mathbf{D} = 0 \quad (73)$$

or

$$\frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = 0$$

For a plane wave given by Eq. (70) the above equation becomes

$$i(k_x D_x + k_y D_y + k_z D_z) = 0$$

or

$$\mathbf{D} \cdot \mathbf{k} = 0 \quad (74)$$

implying that \mathbf{D} is *always* at right angles to \mathbf{k} . Similarly since in a non-magnetic medium $\text{div } \mathbf{H} = 0$,

$$\mathbf{H} \text{ will always be right angles to } \mathbf{k}. \quad (75)$$

Now, in the absence of any currents (i.e., $\mathbf{J} = 0$) Maxwell's curl equations [see Eqs (7) and (8) of Chapter 20] become

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} = i\omega \mathbf{B} = i\omega \mu_0 \mathbf{H} \quad (76)$$

and

$$\nabla \times \mathbf{H} = \frac{\partial \mathbf{D}}{\partial t} = -i\omega \mathbf{D} \quad (77)$$

where we have assumed the medium to be non-magnetic (i.e., $\mathbf{B} = \mu_0 \mathbf{H}$). Now, if

$$\mathbf{E} = \mathbf{E}_0 e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)}$$

then

$$\begin{aligned} (\nabla \times \mathbf{E})_x &= \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} \\ &= (ik_y E_{0z} - ik_z E_{0y}) e^{i(\mathbf{k} \cdot \mathbf{r} - \omega t)} \\ &= i(k_y E_z - k_z E_y) = i(\mathbf{k} \times \mathbf{E})_x \end{aligned}$$

Thus

$$\nabla \times \mathbf{E} = i(\mathbf{k} \times \mathbf{E}) = i\omega \mu_0 \mathbf{H}$$

$$\Rightarrow \mathbf{H} = \frac{1}{\omega \mu_0} (\mathbf{k} \times \mathbf{E}) \quad (78)$$

and

$$\nabla \times \mathbf{H} = i(\mathbf{k} \times \mathbf{H}) = -i\omega \mathbf{D}$$

$$\Rightarrow \mathbf{D} = \frac{1}{\omega} (\mathbf{H} \times \mathbf{k}) \quad (79)$$

Equations (78) and (79) show that

$$\mathbf{H} \text{ is at right angles to } \mathbf{k}, \mathbf{E} \text{ and } \mathbf{D} \quad (80)$$

implying

\mathbf{k}, \mathbf{E} and \mathbf{D} will always be in the same plane.

Further [see Eq. (74)]

$$\mathbf{D} \text{ is at right angles to } \mathbf{k} \quad (81)$$

Substituting for \mathbf{H} in Eq. (79), we get

$$\begin{aligned} \mathbf{D} &= \frac{1}{\omega^2 \mu_0} [(\mathbf{k} \times \mathbf{E}) \times \mathbf{k}] \\ &= \frac{1}{\omega^2 \mu_0} [(\mathbf{k} \cdot \mathbf{k}) \mathbf{E} - (\mathbf{k} \cdot \mathbf{E}) \mathbf{k}] \end{aligned} \quad (82)$$

where we have used the vector identity

$$(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = (\mathbf{A} \cdot \mathbf{C}) \mathbf{B} - (\mathbf{B} \cdot \mathbf{C}) \mathbf{A}$$

Thus

$$\begin{aligned} \mathbf{D} &= \frac{k^2}{\omega^2 \mu_0} [\mathbf{E} - (\hat{\mathbf{k}} \cdot \mathbf{E}) \hat{\mathbf{k}}] \\ &= \frac{n_w^2}{c^2 \mu_0} [\mathbf{E} - (\hat{\mathbf{k}} \cdot \mathbf{E}) \hat{\mathbf{k}}] \end{aligned} \quad (83)$$

where

$$\hat{\mathbf{k}} = \frac{\mathbf{k}}{k} \quad (84)$$

represents the unit vector along \mathbf{k} (see Fig. 19.28). Since

$$D_x = \epsilon_x E_x = \epsilon_0 n_x^2 E_x$$

we have for the x -component of Eq. (83)

$$\frac{\epsilon_0 \mu_0 c^2 n_x^2}{n_w^2} E_x = E_x - \kappa_x (\kappa_x E_x + \kappa_y E_y + \kappa_z E_z)$$

Since $c^2 = 1/(\epsilon_0 \mu_0)$, we have

$$\left(\frac{n_x^2}{n_w^2} - \kappa_x^2 - \kappa_y^2 - \kappa_z^2 \right) E_x + \kappa_x \kappa_y E_y + \kappa_x \kappa_z E_z = 0 \quad (85)$$

where we have used the relation $\kappa_x^2 + \kappa_y^2 + \kappa_z^2 = 1$ (since $\hat{\mathbf{k}}$ is a unit vector). Similarly,

$$\kappa_x \kappa_y E_x + \left(\frac{n_y^2}{n_w^2} - \kappa_x^2 - \kappa_z^2 \right) E_y + \kappa_y \kappa_z E_z = 0 \quad (86)$$

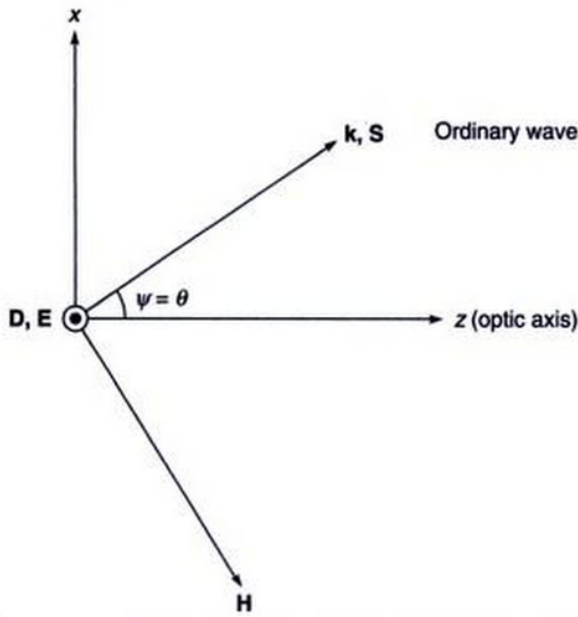


Fig. 19.29 For the ordinary wave (in uniaxial crystals), **D** and **E** vectors are in the *y* direction; **k** and **S** are in the same direction in the *x-z* plane and **H** also lies in the *x-z* plane.

We use Eqs (90) – (92) to obtain

$$\frac{E_z}{E_x} = -\frac{\frac{n_o^2}{n_w^2} - \cos^2 \psi}{\sin \psi \cos \psi} = -\frac{\sin \psi \cos \psi}{\frac{n_e^2}{n_w^2} - \sin^2 \psi}$$

Simple manipulations would give us

$$\frac{1}{n_w^2} = \frac{1}{n_{we}^2} = \frac{\cos^2 \psi}{n_o^2} + \frac{\sin^2 \psi}{n_e^2} \quad (96)$$

where the subscript *e* refers to the fact that the wave refractive index corresponds to the extra-ordinary wave. The corresponding wave velocity would be given by

$$v_{we}^2 = \frac{c^2}{n_{we}^2} = \frac{c^2}{n_o^2} \cos^2 \psi + \frac{c^2}{n_e^2} \sin^2 \psi \quad (97)$$

Since the wave velocity is dependent on the direction of the wave, it is referred to as the extra-ordinary wave and hence the subscript *e*. Of course, for the extra-ordinary wave, we must have

$$D_y = \epsilon_y E_y = 0$$

From the above equation and Eq.(79), it follows that the displacement vector **D** of the wave is normal to the *y*-axis and also to **k** implying that the displacement vector **D** associated with the extraordinary wave lies in the plane containing the propagation vector **k** and the optic axis and is normal to **k** (see Fig. 19.30). This was the recipe given

through Eq. (42). Figure 19.30 also shows the Poynting vector **S** ($= \mathbf{E} \times \mathbf{H}$) which represents the direction of energy propagation (i.e., the direction of the *e*-ray). The small dashes on the extraordinary ray in Figs. 19.18(a) and (b) represent the directions of the **D** vector. Let ϕ and θ represent the angles that the **S** vector makes with the **k** vector and the optic axis respectively (see Fig. 19.29). In order to determine the angle ϕ we note that

$$\frac{\epsilon_z E_z}{\epsilon_x E_x} = \frac{D_z}{D_x} = -\tan \psi$$

and since

$$\frac{E_z}{E_x} = -\tan (\phi + \psi) \quad (98)$$

we get

$$\frac{n_e^2}{n_o^2} \tan (\phi + \psi) = \tan \psi$$

or,

$$\phi = \tan^{-1} \left[\frac{n_o^2}{n_e^2} \tan \psi \right] - \psi$$

Obviously, for negative crystals $n_o > n_e$ and ϕ will be positive implying that ray direction is further away from the optic axis as shown in Fig. 19.30.

Conversely, for positive crystals $n_o < n_e$ and ϕ will be negative implying that the ray direction will be towards the optic axis.

Example 19.3 We consider calcite for which (at $\lambda = 5893 \text{ \AA}$ and 18°C)

$$n_o = 1.65835, n_e = 1.48640$$

If we consider **k** making an angle of 30° to the optic axis, then $\psi = 30^\circ$ and elementary calculations give us $\phi = 5.7^\circ$

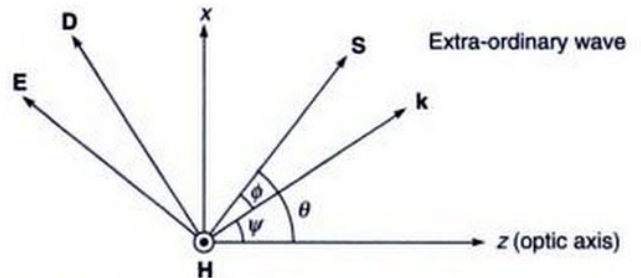


Fig. 19.30 For the extra-ordinary wave (in uniaxial crystals), **E**, **D**, **S** and **k** vectors would lie in the *x-z* plane and **H** will be in the *y* direction. **S** is at right angles to **E** and **H**; **D** is at right angles to **k** and **H**.

19.13 RAY VELOCITY AND RAY REFRACTIVE INDEX

The direction of energy propagation (or the ray propagation) is along the Poynting vector \mathbf{S} which is given by

$$\mathbf{S} = \mathbf{E} \times \mathbf{H} \quad (99)$$

Thus, since the plane containing the vectors \mathbf{k} , \mathbf{E} and \mathbf{D} is normal to \mathbf{H} , the Poynting vector \mathbf{S} will also lie in the plane containing the vectors \mathbf{k} , \mathbf{E} and \mathbf{D} (see Figs 19.29 and 19.30). For the extra-ordinary wave, the direction of the propagation of the wave ($\hat{\mathbf{k}}$) is not along the direction of energy propagation ($\hat{\mathbf{s}}$), where $\hat{\mathbf{s}}$ is the unit vector along \mathbf{S} . The ray velocity (or the energy transmission velocity) v_r is defined as

$$v_r = \frac{S}{u} \quad (100)$$

where u is the energy density. Now,

$$\begin{aligned} u &= \frac{1}{2} (\mathbf{D} \cdot \mathbf{E} + \mathbf{B} \cdot \mathbf{H}) \\ &= \frac{1}{2} (\mathbf{D} \cdot \mathbf{E} + \mu_0 \mathbf{H} \cdot \mathbf{H}) \end{aligned} \quad (101)$$

[see Eq. (58) of Chapter 20]. Substituting for \mathbf{H} and \mathbf{D} from Eqs (78) and (79), we obtain

$$\begin{aligned} u &= \frac{1}{2\omega} [(\mathbf{H} \times \mathbf{k}) \cdot \mathbf{E} + (\mathbf{k} \times \mathbf{E}) \cdot \mathbf{H}] \\ &= \frac{1}{2\omega} [\mathbf{k} \cdot (\mathbf{E} \times \mathbf{H}) + \mathbf{k} \cdot (\mathbf{E} \times \mathbf{H})] \\ &= \frac{1}{\omega} \mathbf{k} \cdot \mathbf{S} \end{aligned} \quad (102)$$

Thus Eq. (100) becomes

$$v_r = \frac{\omega S}{\mathbf{k} \cdot \mathbf{S}} = \frac{\omega}{k \cos \phi} = \frac{v_w}{\cos \phi} \quad (103)$$

where ϕ is the angle between $\hat{\mathbf{k}}$ and $\hat{\mathbf{s}}$ (see Fig. 19.30). The ray refractive index n_r is defined as

$$n_r = \frac{c}{v_r} = \frac{c}{v_w} \cos \phi = n_w \cos \phi \quad (104)$$

In order to express \mathbf{E} in terms of \mathbf{D} , we refer to Fig. 19.30 and write

$$\mathbf{D} = (\mathbf{D} \cdot \hat{\mathbf{e}}) \hat{\mathbf{e}} + (\mathbf{D} \cdot \hat{\mathbf{s}}) \hat{\mathbf{s}}$$

where $\hat{\mathbf{e}}$ is a unit vector along the direction of the electric field \mathbf{E} . Thus

$$\mathbf{D} - (\mathbf{D} \cdot \hat{\mathbf{s}}) \hat{\mathbf{s}} = (\mathbf{D} \cdot \hat{\mathbf{e}}) \hat{\mathbf{e}} = (D \cos \phi) \frac{\mathbf{E}}{E} \quad (105)$$

Similarly,

$$\mathbf{E} = (\mathbf{E} \cdot \hat{\mathbf{d}}) \hat{\mathbf{d}} + (\mathbf{E} \cdot \hat{\mathbf{k}}) \hat{\mathbf{k}} \quad (106)$$

where $\hat{\mathbf{d}}$ represents a unit vector along the displacement vector \mathbf{D} (see Fig 19.30). If we now substitute for $\mathbf{E} - (\mathbf{E} \cdot \hat{\mathbf{k}}) \hat{\mathbf{k}}$ in Eq. (83), we would get

$$\mathbf{D} = \frac{n_w^2}{\mu_0 c^2} (\mathbf{E} \cdot \hat{\mathbf{d}}) \hat{\mathbf{d}}$$

or

$$D = \frac{n_w^2}{\mu_0 c^2} E \cos \phi \quad (107)$$

Substituting in Eq. (105), we get

$$\mathbf{D} - (\mathbf{D} \cdot \hat{\mathbf{s}}) \hat{\mathbf{s}} = \frac{n_w^2}{\mu_0 c^2} \cos^2 \phi \mathbf{E} = \frac{n_r^2}{\mu_0 c^2} \mathbf{E}$$

where, in the last step, we have used Eq. (104). Taking the x component of the above equation (where x represents the direction of one of the principal axes), we obtain

$$\begin{aligned} D_x - (D_x s_x + D_y s_y + D_z s_z) s_x \\ = \frac{n_r^2}{\mu_0 c^2} E_x = \frac{n_r^2}{\mu_0 c^2 \epsilon_x} D_x \end{aligned}$$

If we use the relations

$$n_x^2 = \frac{\epsilon_x}{\epsilon_0}, c^2 = \frac{1}{\epsilon_0 \mu_0} \quad \text{and} \quad s_x^2 + s_y^2 + s_z^2 = 1$$

we would get

$$\left(\frac{n_r^2}{n_x^2} - s_y^2 - s_z^2 \right) D_x + s_x s_y D_y + s_x s_z D_z = 0 \quad (108)$$

Similarly

$$s_x s_y D_x + \left(\frac{n_r^2}{n_y^2} - s_x^2 - s_z^2 \right) D_y + s_z s_y D_z = 0 \quad (109)$$

$$s_x s_z D_x + s_z s_y D_y + \left(\frac{n_r^2}{n_z^2} - s_x^2 - s_y^2 \right) D_z = 0 \quad (110)$$

As in the previous section, the above set of equations form a set of three homogenous equations. For non-trivial solutions, we must have

$$\begin{vmatrix} \frac{n_r^2}{n_x^2} - s_y^2 - s_z^2 & s_x s_y & s_x s_z \\ s_x s_y & \frac{n_r^2}{n_y^2} - s_x^2 - s_z^2 & s_z s_y \\ s_x s_z & s_z s_y & \frac{n_r^2}{n_z^2} - s_x^2 - s_y^2 \end{vmatrix} = 0 \quad (111)$$

which represents a LP beam with its E oscillating at right angles to the direction of the incident polarization. It may be noted that two QWP put together makes a HWP. Let us next consider the incidence of a LEP given by

$$\begin{aligned} E_y &= \frac{E_0}{2} \cos(kx - \omega t) \\ E_z &= \frac{\sqrt{3}}{2} E_0 \sin(kx - \omega t) \\ &= \frac{\sqrt{3}}{2} E_0 \operatorname{Re} e^{i(kx - \omega t - \frac{\pi}{2})} \end{aligned}$$

Thus the beam coming out of the QWP will be given by

$$\begin{pmatrix} E'_y \\ E'_z \end{pmatrix} = \begin{pmatrix} i & 0 \\ 0 & 1 \end{pmatrix} E_0 \begin{pmatrix} 1/2 \\ -i\sqrt{3}/2 \end{pmatrix}$$

which is LP with its E making an angle of 120° with the z -axis.

The use of Jones matrices makes it very straightforward to consider more complicated cases like two QWP with their axes at an angle.

19.15 FARADAY ROTATION

Consider a linearly polarized light propagating through a medium. If a magnetic field is applied along the direction of propagation of the polarized wave, then the plane of polarization gets rotated—this rotation is usually referred to as *Faraday rotation* after the famous physicist Michael Faraday who discovered this phenomenon in 1845. In the presence of a (longitudinal) magnetic field, the modes of propagation are the left circularly polarized (LCP) wave and the right circularly polarized (RCP) wave. Thus the situation is somewhat similar to the phenomenon of optical activity discussed in Sec. 19.8. The angle θ by which the plane of polarization rotates is given by the empirical formula

$$\theta = VHl$$

where H is the applied magnetic field, l is the length of the medium and V is called the Verdet constant. For silica $V \approx 2.64 \times 10^{-4}$ deg/Ampere $\approx 4.6 \times 10^{-6}$ radians/Ampere.

The Faraday rotation has a very important application in measuring large currents using single mode optical fibers. We consider a large length of a single mode fiber wound in many turns in the form of a loop around a current carrying conductor (see Fig 19.31). If a current I is passing through the conductor then by Ampere's law

$$\int \mathbf{H} \cdot d\mathbf{l} = NI$$

where N represents the number of loops of the fiber around the conductor. Thus if a linearly polarized light is incident

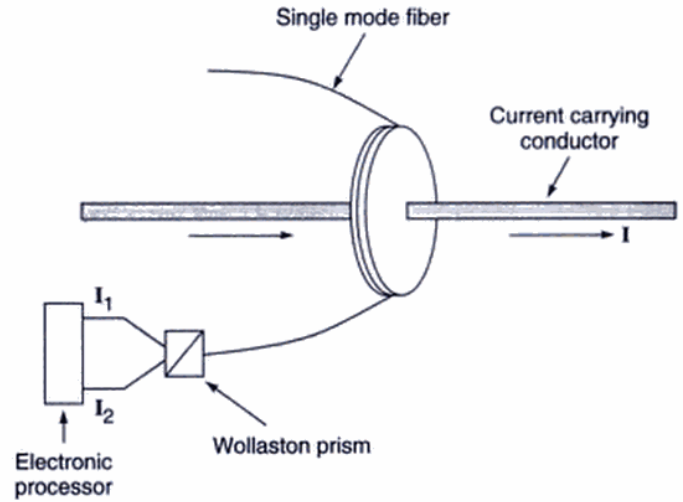


Fig. 19.31 A single mode fiber wound helically around a current carrying conductor. The rotation of the plane of polarization is detected by passing the light through a Wollaston prism and then an electronic processor.

on the fiber, then its plane of polarization will get rotated by the angle

$$\theta = VNI$$

The rotation θ does not depend on the shape of the loop. As an example, for $I = 200$ Amperes and $N = 50$, $\theta \approx 0.26$ degree. The light from the fiber is allowed to fall on a Wollaston prism and the outputs are measured separately; the Faraday rotation θ is given by

$$\theta = \text{constant} \frac{I_1 - I_2}{I_1 + I_2}$$

where I_1 , and I_2 are the currents in the electronic processor due to the two beams coming out of the Wollaston prism. Figure 19.32 shows an actual variation of the output with the current passing through the conductor. Such a set up can be used to measure very high currents (~ 10000 Amperes).

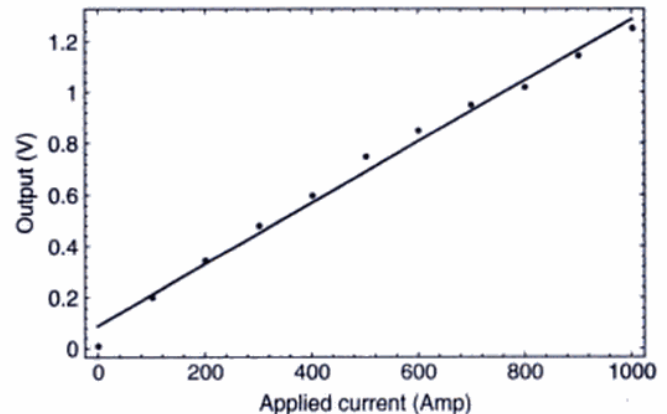


Fig. 19.32 A typical variation of the output signal with current [Figure kindly provided by Dr. Parthasarathi Palai].

19.16 THEORY OF OPTICAL ACTIVITY

As mentioned earlier, in an isotropic dielectric, the \mathbf{D} vector is in the same direction as \mathbf{E} and we have

$$\mathbf{D} = \epsilon \mathbf{E} = \epsilon_0 n^2 \mathbf{E} \quad (129)$$

where $\epsilon_0 (= 8.854 \times 10^{-12}$ MKS units) is the permittivity of free space and $n (= \sqrt{\epsilon/\epsilon_0})$ is the refractive index of the medium. Now, if we dissolve cane sugar in water, the medium is still isotropic; however, because of the spiral like structure of sugar molecules, the relation between \mathbf{D} and \mathbf{E} is given by the following relation

$$\begin{aligned} \mathbf{D} &= \epsilon_0 n^2 \mathbf{E} + ig \hat{\mathbf{k}} \times \mathbf{E} \\ &= \epsilon_0 n^2 [\mathbf{E} + i\alpha \hat{\mathbf{k}} \times \mathbf{E}] \end{aligned} \quad (130)$$

where

$$\alpha = \frac{g}{\epsilon_0 n^2}$$

and $\hat{\mathbf{k}}$ is the unit vector along the direction of propagation of the wave. The parameter α can be positive or negative but it is usually an extremely small number ($\ll 1$). Without any loss of the generality, we may assume propagation along the z -axis so that $\kappa_x = \kappa_y = 0$ and $\kappa_z = 1$ giving

$$\hat{\mathbf{k}} \times \mathbf{E} = \begin{pmatrix} \hat{x} & \hat{y} & \hat{z} \\ 0 & 0 & 1 \\ E_x & E_y & E_z \end{pmatrix} = -\hat{x} E_y + \hat{y} E_x$$

Thus

$$\begin{pmatrix} D_x \\ D_y \\ D_z \end{pmatrix} = \begin{pmatrix} \epsilon_0 n^2 & -ig & 0 \\ ig & \epsilon_0 n^2 & 0 \\ 0 & 0 & \epsilon_0 n^2 \end{pmatrix} \begin{pmatrix} E_x \\ E_y \\ E_z \end{pmatrix} \quad (131)$$

The ϵ matrix is still Hermitian but there is a "small" off diagonal imaginary element. The presence of these off diagonal terms give rise to optical activity. We rewrite Eq. (83)

$$\frac{n_w^2}{c^2 \mu_0} [\mathbf{E} - (\hat{\mathbf{k}} \cdot \mathbf{E}) \hat{\mathbf{k}}] = \mathbf{D}$$

We write the x and y components of the above equation and since $\kappa_x = 0 = \kappa_y$ and $\kappa_z = 1$, we get

$$\frac{n_w^2}{c^2 \mu_0} E_x = D_x = \epsilon_0 n^2 E_x - ig E_y$$

and

$$\frac{n_w^2}{c^2 \mu_0} E_y = D_y = ig E_x + \epsilon_0 n^2 E_y$$

Thus

$$\left(\frac{n_w^2}{n^2} - 1 \right) E_x = -i\alpha E_y$$

and

$$\left(\frac{n_w^2}{n^2} - 1 \right) E_y = i\alpha E_x$$

where we have used the fact that $c = 1/\sqrt{\epsilon_0 \mu_0}$. For non-trivial solutions

$$\left(\frac{n_w^2}{n^2} - 1 \right)^2 = \alpha^2$$

giving

$$n_w = n\sqrt{1 \pm \alpha} \quad (132)$$

and

$$E_y = \pm i E_x \quad (133)$$

We write the two solutions as $n_r (= n\sqrt{1+\alpha})$ and $n_l (= n\sqrt{1-\alpha})$; the corresponding propagation constants will be given by

$$k = k_r = \frac{\omega}{c} n_r = \frac{\omega}{c} n\sqrt{1+\alpha} \quad (134)$$

and

$$k = k_l = \frac{\omega}{c} n_l = \frac{\omega}{c} n\sqrt{1-\alpha} \quad (135)$$

For $n_w = n_r$, if

$$E_x = E_0 e^{i(k_r z - \omega t)}$$

then

$$E_y = +i E_x = E_0 e^{i(k_r z - \omega t + \frac{\pi}{2})}$$

which would represent an RCP (Right Circularly Polarized) wave and hence the subscript r . Similarly, For $n_w = n_l$, if

$$E_x = E_0 e^{i(k_l z - \omega t)}$$

then

$$E_y = -i E_x = E_0 e^{i(k_l z - \omega t - \frac{\pi}{2})}$$

which would represent an LCP (Left Circularly Polarized) wave and hence the subscript l . The RCP and LCP waves are the two "modes" of the "optically active" substance and for an arbitrary incident state of polarization, we must write it as a superposition of the two modes and study the independent propagation of the two modes. Now,

$$\begin{aligned} n_r - n_l &= n[\sqrt{1+\alpha} - \sqrt{1-\alpha}] \\ &\approx n\alpha \end{aligned} \quad (136)$$

and

$$\frac{d^2 y}{dt^2} + \omega_0^2 y - \frac{qB_0}{m} \frac{dx}{dt} = \mp i \frac{q}{m} E_0 e^{i(kz - \omega t)} \quad (146)$$

where the upper and lower signs correspond to RCP and LCP respectively. Writing

$$x = x_0 e^{i(kz - \omega t)} \quad \text{and} \quad y = y_0 e^{i(kz - \omega t)}$$

we get

$$\left. \begin{aligned} (\omega^2 - \omega_0^2) x_0 + i\omega_c \omega y_0 &= + \frac{q}{m} E_0 \\ (\omega^2 - \omega_0^2) y_0 - i\omega_c \omega x_0 &= \pm i \frac{q}{m} E_0 \end{aligned} \right\} \begin{array}{l} \times (\omega^2 - \omega_0^2) \\ \times -i\omega_c \omega \end{array} \quad (147, 148)$$

where

$$\omega_c = \frac{qB_0}{m}$$

is the electron cyclotron frequency. If we multiply Eq.(147) by $(\omega^2 - \omega_0^2)$ and Eq.(148) by $-i\omega_c \omega$ and add the two equations we would get

$$[(\omega^2 - \omega_0^2)^2 - \omega_c^2 \omega^2] x_0 = \frac{q}{m} E_0 [(\omega^2 - \omega_0^2) \pm \omega_c \omega]$$

giving

$$x_0 = \frac{qE_0}{m[(\omega^2 - \omega_0^2) \mp \omega_c \omega]}$$

Similarly

$$y_0 = \pm i \frac{qE_0}{m[(\omega^2 - \omega_0^2) \mp \omega_c \omega]} = \pm i x_0$$

Thus the polarization is given by

$$\begin{aligned} \mathbf{P} &= -Nq\mathbf{r} \\ &= -Nq \cdot \frac{qE_0(\hat{x} \pm i\hat{y})}{m[(\omega^2 - \omega_0^2) \mp \omega_c \omega]} e^{i(kz - \omega t)} \\ &= \chi \mathbf{E}_{\pm} \end{aligned}$$

where the susceptibility χ is given by

$$\chi = \frac{Nq^2}{m} \cdot \frac{1}{[(\omega_0^2 - \omega^2) \pm \omega_c \omega]}$$

Thus the modes are circularly polarized and the corresponding refractive indices are given by [cf. Eq. (84) of Chapter 6]

$$n_{\pm}^2 = 1 + \frac{Nq^2}{m\epsilon_0} \cdot \frac{1}{[(\omega_0^2 - \omega^2) \pm \omega_c \omega]}$$

where the upper and lower signs correspond to RCP and LCP respectively.

SUMMARY

- For an electromagnetic wave propagating in the z -direction, let x and y components of the electric field be given by

$$E_x = a_1 \cos(kz - \omega t)$$

$$E_y = a_2 \cos(kz - \omega t + \theta)$$

- (i) If $a_2 = 0$, we have a x -polarized wave. If $a_1 = 0$, we have a y -polarized wave. For $\theta = n\pi$; $n = 0, \pm 1, \pm 2, \dots$ we again have a linearly polarized wave with the electric vector making an angle with the x -axis—this angle is either $+\tan^{-1}\left(\frac{a_2}{a_1}\right)$ or $-\tan^{-1}\left(\frac{a_2}{a_1}\right)$.

- (ii) If $a_1 = a_2$ and $\theta = \frac{\pi}{2}, \frac{5\pi}{2}, \frac{9\pi}{2}, \dots$ we have a RCP (right circularly polarized) wave; for $\theta = \frac{3\pi}{2}, \frac{7\pi}{2}, \dots$ we have a LCP (left circularly polarized) wave.
- (iii) In general, we have either a LEP (left elliptically polarized) wave or a REP (right elliptically polarized) wave.

- Linearly polarized light can be produced by various methods: e.g.,

- (a) by allowing an unpolarized light to pass through a polaroid
- (b) by allowing an unpolarized light to fall on a dielectric surface at the Brewster angle

$$\theta_p = \left(\tan^{-1} \frac{n_2}{n_1} \right).$$

- (c) by passing through a Nicol prism

- If an unpolarized plane wave is incident on an uniaxial crystal, the plane wave will split into two plane waves. One is referred to as the ordinary wave (usually abbreviated as the o -wave) and the other is referred to as the extra-ordinary wave (usually abbreviated as the e - o -wave). For both waves, the space and time dependence of the vectors \mathbf{E} , \mathbf{D} , \mathbf{B} and \mathbf{H} can be assumed to be of the form $e^{i(\mathbf{k}\cdot\mathbf{r} - \omega t)}$ where \mathbf{k} denotes the propagation vector and represents the direction normal to the phase fronts. In general, \mathbf{k} vector for the o - and e -waves will be different. Further,

- (a) Both ordinary and extra-ordinary waves are linearly polarized.
- (b) $\mathbf{D} \cdot \mathbf{k} = 0$ for both o - and e -waves
- (c) For the o -wave, the \mathbf{D} vector is at right angles to the optic axis as well as to \mathbf{k} .

- 19.7** Show that the angle between the vectors \mathbf{D} and \mathbf{E} is the same as between the Poynting vector \mathbf{S} and the propagation vector \mathbf{k} .
- 19.8** Consider the propagation of an extra-ordinary wave through a KDP crystal. If the wave vector is at an angle of 45° to the optic axis, calculate the angle between \mathbf{S} and \mathbf{k} . Repeat the calculation for LiNbO_3 . The values of n_o and n_e for KDP and LiNbO_3 are given in Table 19.1.

[Ans. 1.56° and 2.25°]

- 19.9** Prove that when the angle of incidence corresponds to the Brewster angle, the reflected and refracted rays are at right angles to each other.
- 19.10** (a) Consider two crossed polaroids placed in the path of an unpolarized beam of intensity I_0 (see Fig. 19.6). If we place a third polaroid in between the two then, in general, some light will be transmitted through. Explain this phenomenon.
- (b) Assuming the pass axis of the third polaroid to be at 45° to the pass axis of either of the polaroids, calculate the intensity of the transmitted beam. Assume that all the polaroids are perfect.

[Ans. $1/8 I_0$]

- 19.11** A quarter-wave plate is rotated between two crossed polaroids. If an unpolarized beam is incident on the first polaroid, discuss the variation of intensity of the emergent beam as the quarter-wave plate is rotated. What will happen if we have a half-wave instead of a quarter-wave plate?
- 19.12** In Problem 19.11, if the optic axis of the quarter-wave plate makes an angle of 45° with the pass axis of either polaroid, show that only a quarter of the incident intensity will be transmitted. If the quarter-wave plate is replaced by a half-wave plate, show that half of the incident intensity will be transmitted through.
- 19.13** For calcite the values of n_o and n_e for $\lambda_0 = 4046\text{\AA}$ are 1.68134 and 1.49694 respectively; corresponding to $\lambda_0 = 7065\text{\AA}$ the values are 1.65207 and 1.48359 respectively. We have a calcite quarter-wave plate corresponding to $\lambda_0 = 4046\text{\AA}$. A left-circularly polarized beam of $\lambda_0 = 7065\text{\AA}$ is incident on this plate. Obtain the state of polarization of the emergent beam.
- 19.14** A HWP (half wave plate) is introduced between two crossed polaroids P_1 and P_2 . The optic axis makes an angle 15° with the pass axis of P_1 as shown in Fig. 19.33(a) and (b). If an unpolarized beam of intensity I_0 is normally incident on P_1 and if I_1 , I_2 , and

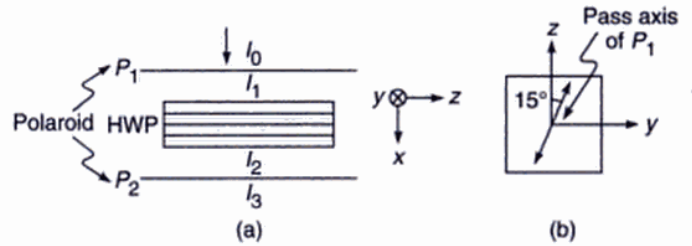


Fig. 19.33

I_3 are the intensities after P_1 , after HWP and after P_2 respectively then calculate I_1/I_0 , I_2/I_0 and I_3/I_0 .

[Ans. $1/2$, $1/2$, $1/8$]

- 19.15** Two prisms of calcite ($n_o > n_e$) are cemented together as shown in Fig. 19.34, so as to form a cube. Lines and dots show the direction of the optic axis. A beam of unpolarized light is incident normally from region I. Assume the angle of the prism to be 12° . Determine the path of rays in regions II, III & IV indicating the direction of vibrations (i.e., the direction of \mathbf{D}).

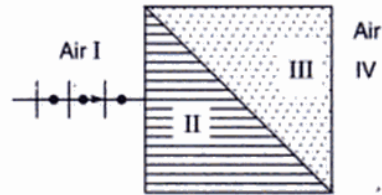


Fig. 19.34

- 19.16** A $\lambda/6$ plate is introduced in between the two crossed polarizers in such a way that the optic axis of the $\lambda/6$ plate makes an angle of 45° with the pass axis of the first polarizer (see Fig. 19.35). Consider an unpolarized beam of intensity I_0 to be incident normally on the polarizer. Assume the optic axis to be along the z -axis and the propagation along the x -axis. Write the y and z components of the electric fields (and the corresponding total intensities) after passing through (i) P_1 (ii) $\lambda/6$ plate and (iii) P_2 .

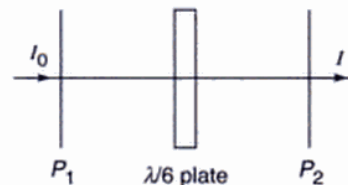


Fig. 19.35

- 19.17** A beam of light is passed through a polarizer. If the polarizer is rotated with the beam as an axis, the intensity I of the emergent beam does not vary. What

are the possible states of polarization of the incident beam? How to ascertain its state of polarization with the help of the given polarizer and a QWP?

- 19.18** Consider a Wollaston prism consisting of two similar prisms of calcite ($n_o = 1.66$ and $n_e = 1.49$) as shown in Fig. 19.26, with angle of prism now equal to 25° . Calculate the angular divergence of the two emerging beams.

- 19.19** (a) Consider a plane wave incident normally on a calcite crystal with its optic axis making an angle of 20° with the normal [see Fig. 19.18(a)]. Thus $\psi = 20^\circ$. Calculate the angle that the Poynting vector will make with the normal to the surface. Assume $n_o \approx 1.66$ and $n_e \approx 1.49$.
 (b) In the above problem assume the crystal to be quartz with $n_o \approx 1.544$ and $n_e \approx 1.553$.

[Ans. (a) 4.31°]

- 19.20** Consider the incidence of the following REP beam on a sugar solution at $z = 0$:

$$E_x = 5 \cos \alpha x; E_y = 4 \sin \alpha x$$

with $\lambda = 6328 \text{ \AA}$. Assume

$$n_l - n_r = 10^{-5} \text{ and } n_l = 4/3$$

Study the evolution of the SOP of the beam.

- 19.21** Consider the incidence of the above REP beam on an elliptic core fiber with

$$\frac{\beta_x}{k_0} \approx 1.506845 \text{ and } \frac{\beta_y}{k_0} \approx 1.507716$$

Calculate the SOP at $z = 0.25 L_b, 0.5 L_b, 0.75 L_b$ and L_b .

- 19.22** When the optic axis lies on the surface of the crystal and in the plane of incidence, show (by geometrical considerations) that the angles of refraction of the ordinary and the extra-ordinary rays (which we denote by r_o and r_e respectively) are related through the following equation:

$$\frac{\tan r_o}{\tan r_e} = \frac{n_o}{n_e}$$

REFERENCES AND SUGGESTED READINGS

1. W. A. Shurcliff and S. S. Ballard, *Polarized Light*, Van Nostrand, Princeton, New Jersey, USA (1964).
2. G. R. Bird and M. P. Parrish, *The wire grid as a near infrared polarizer*, J. Opt. Soc. Am (1960), **50**, 886.
3. M. Alonso and E. J. Finn, *Physics*, Addison-Wesley, Reading, Massachusetts, USA (1970).
4. R. P. Feynman, R. B. Leighton and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley, Reading, Massachusetts, USA (1963).
5. M. Born and E. Wolf, *Principles of Optics*, Pergamon Press, Oxford, England (1970).
6. A. K. Ghatak and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, Cambridge, UK (1989). [Reprinted by Foundation Books, New Delhi]
7. F. A. Jenkins and H. E. White, *Fundamentals of Optics*, McGraw-Hill, New York, USA (1976).
8. *Polarized Light: Selected Reprints*, American Institute of Physics, New York, USA (1963).
9. S. Chandrasekhar, *Simple model for optical activity*, Amer. J. Phys., **24**, 503 (1956).
10. P. Gay, *An Introduction to Crystal Optics*, Longmans Green and Co, London, England (1967).
11. T. H. Waterman, *Polarized light and animal navigation*, Scien. Amer (July, 1955).
12. E. A. Wood, *Crystals and Light*, Van Nostrand Momentum Book No. 5, Van Nostrand, Princeton, New Jersey, USA (1964).
13. L.B. Jeunhomme, *Single mode fiber optics*, Marcel Dekker, New York (1983).

Chapter 20

Electromagnetic Waves

*And God said, "Let there be light";
and there was light.*

—Genesis

*Nature, and Nature's laws, lay hid in night;
God said, Let Newton be! and all was light.*

—Alexander Pope

*Maxwell could say, when he was finished with his discovery, 'Let there be electricity and magnetism,
and there is light'*

—Feynman

20.1 MAXWELL'S EQUATIONS

All electromagnetic phenomena can be said to follow from Maxwell's equations. These equations are based on experimental observations and are given by the following equations:

$$\frac{\partial D_x}{\partial x} + \frac{\partial D_y}{\partial y} + \frac{\partial D_z}{\partial z} = \rho \quad (1)$$

$$\frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0 \quad (2)$$

$$\left. \begin{aligned} \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} &= -\frac{\partial B_x}{\partial t} \\ \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= -\frac{\partial B_y}{\partial t} \\ \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} &= -\frac{\partial B_z}{\partial t} \end{aligned} \right\} \quad (3)$$

$$\left. \begin{aligned} \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} &= J_x + \frac{\partial D_x}{\partial t} \\ \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} &= J_y + \frac{\partial D_y}{\partial t} \\ \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= J_z + \frac{\partial D_z}{\partial t} \end{aligned} \right\} \quad (4)$$

where ρ represents the charge density and \mathbf{J} the current density; \mathbf{E} , \mathbf{D} , \mathbf{B} and \mathbf{H} represent the electric field, electric displacement, magnetic induction and magnetic field respectively. The physical implications of each equation will be discussed in Sec. 20.9; we may note here that using vector notation the above equations can be written in the following compact form:

$$\text{div } \mathbf{D} = \rho \quad (5)$$

$$\text{div } \mathbf{B} = 0 \quad (6)$$

$$\text{curl } \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \quad (7)$$

$$\text{curl } \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (8)$$

The above equations can be solved only if the 'constitutive relations' are known which relate \mathbf{D} to \mathbf{E} , \mathbf{B} to \mathbf{H} and \mathbf{J} to \mathbf{E} . For a linear, isotropic and homogeneous medium the 'constitutive relations' are given by the following equations*

$$\mathbf{D} = \epsilon \mathbf{E} \quad (9)$$

$$\mathbf{B} = \mu \mathbf{H} \quad (10)$$

and

$$\mathbf{J} = \sigma \mathbf{E} \quad (11)$$

where ϵ , μ and σ denote respectively the dielectric permittivity, magnetic permeability and conductivity of the medium respectively.

*It should be mentioned that the 'constitutive relations' depend on the properties of the medium, field strengths, etc. For example, for an anisotropic medium ϵ is a tensor of second rank (see Sec. 19.8); for high field strengths ϵ may itself depend on \mathbf{E} (see Sec. 16.10); etc.

20.2 PLANE WAVES IN A DIELECTRIC

For a non-charged current-free dielectric, we will have

$$\rho = 0 \quad (12)$$

$$\text{and} \quad \mathbf{J} = 0 \quad (13)$$

Further, if we use the 'constitutive relations' [Eqs (9)–(11)], then Eqs. (1)–(4) become

$$\frac{\partial E_x}{\partial x} + \frac{\partial E_y}{\partial y} + \frac{\partial E_z}{\partial z} = 0 \quad (14)$$

$$\frac{\partial H_x}{\partial x} + \frac{\partial H_y}{\partial y} + \frac{\partial H_z}{\partial z} = 0 \quad (15)$$

$$\left. \begin{aligned} \frac{\partial E_z}{\partial y} - \frac{\partial E_y}{\partial z} &= -\mu \frac{\partial H_x}{\partial t} \\ \frac{\partial E_x}{\partial z} - \frac{\partial E_z}{\partial x} &= -\mu \frac{\partial H_y}{\partial t} \\ \frac{\partial E_y}{\partial x} - \frac{\partial E_x}{\partial y} &= -\mu \frac{\partial H_z}{\partial t} \end{aligned} \right\} \quad (16)$$

and

$$\left. \begin{aligned} \frac{\partial H_z}{\partial y} - \frac{\partial H_y}{\partial z} &= \epsilon \frac{\partial E_x}{\partial t} \\ \frac{\partial H_x}{\partial z} - \frac{\partial H_z}{\partial x} &= \epsilon \frac{\partial E_y}{\partial t} \\ \frac{\partial H_y}{\partial x} - \frac{\partial H_x}{\partial y} &= \epsilon \frac{\partial E_z}{\partial t} \end{aligned} \right\} \quad (17)$$

In Sec. 20.3, we will (using the above equations) derive the wave equation; however, in this section we will assume the existence of electromagnetic waves and study the properties of plane waves. For plane waves propagating in the $+z$ -direction, the electric and magnetic fields can be written in the form:

$$\mathbf{E} = \mathbf{E}_0 \exp [i(kz - \omega t)] \quad (18)$$

and

$$\mathbf{H} = \mathbf{H}_0 \exp [i(kz - \omega t)] \quad (19)$$

where \mathbf{E}_0 and \mathbf{H}_0 are space and time independent vectors; but may, in general, be complex. Since \mathbf{E} and \mathbf{H} are independent of x and y , Eqs (14) and (15) give

$$\left. \begin{aligned} \frac{\partial E_z}{\partial z} &= 0 \\ \frac{\partial H_z}{\partial z} &= 0 \end{aligned} \right\} \quad (20)$$

Since we are interested in plane-wave solutions of the type given by Eqs (18) and (19), we must have*

$$E_z = 0 \quad \text{and} \quad H_z = 0 \quad (21)$$

Thus the longitudinal components vanish and hence the waves are of transverse type. Without any loss of generality, we may choose the x -axis to be along \mathbf{E} ; thus we may assume

$$E_y = 0 \quad (22)$$

Using Eqs (18), (19), (21) and (22), Eqs (16) and (17) simplify to

$$0 = i\omega\mu H_{0x} \quad (23)$$

$$ikE_{0x} = i\omega\mu H_{0y} \quad (24)$$

$$\text{and} \quad -ikH_{0y} = -i\omega\epsilon E_{0x} \quad (25)$$

$$ikH_{0x} = 0 \quad (26)$$

where

$$\left. \begin{aligned} \mathbf{E}_0 &= \hat{x}E_{0x} + \hat{y}E_{0y} + \hat{z}E_{0z} \\ \mathbf{H}_0 &= \hat{x}H_{0x} + \hat{y}H_{0y} + \hat{z}H_{0z} \end{aligned} \right\} \quad (27)$$

i.e. E_{0x} , E_{0y} and E_{0z} are the x , y and z -components of \mathbf{E}_0 , etc. Eqs. (23)–(26) give us and

$$\left. \begin{aligned} H_{0x} &= 0 \\ \frac{H_{0y}}{E_{0x}} &= \frac{k}{\omega\mu} = \frac{\omega\epsilon}{k} \end{aligned} \right\} \quad (28)$$

Thus, the plane waves as represented by Eqs (18) and (19) indeed satisfy Maxwell's equations with

$$k^2 = \omega^2\epsilon\mu \quad (29)$$

The waves are transverse in nature and if we assume the electric vector to be along the x -axis, then the magnetic vector will be along the y -axis so that we may write

$$\left. \begin{aligned} \mathbf{E} &= \hat{x}E_0 e^{i(kz - \omega t)} \\ \mathbf{H} &= \hat{y}H_0 e^{i(kz - \omega t)} \end{aligned} \right\} \quad (30)$$

$$\text{with} \quad H_0 = \frac{k}{\omega\mu} E_0 \quad (31)$$

the magnetic induction would, therefore, be given by

$$\mathbf{B} = \hat{y}B_0 e^{i(kz - \omega t)} \quad (32)$$

$$\text{where} \quad B_0 = \frac{k}{\omega} E_0 = \sqrt{\epsilon\mu} E_0 = \frac{1}{v} E_0 \quad (33)$$

$v \left(= \frac{1}{\epsilon\mu} \right)$ represents the speed of propagation of the electromagnetic waves (see Eq. (35)). In free-space

$$B_0 = \frac{1}{c} E_0 \quad (34)$$

*Equation (21) does not represent the unique solution of Eq. (20). For example, we could have chosen $E_z = \text{constant}$; but such a field would not have represented waves.

From the above equations we may draw the following inferences for plane waves:

- (i) Both \mathbf{E} and \mathbf{H} are at right angles to the direction of propagation. Thus the waves are transverse (see Fig. 20.1).
- (ii) The vectors \mathbf{E} and \mathbf{H} are at right angles to each other; thus if the direction of propagation is along the z -axis and if \mathbf{E} is assumed to point in the x -direction then \mathbf{H} will point in the y -direction (see Fig. 20.1).



Fig. 20.1 If a plane wave is propagating in the z -direction (which is coming out of the paper) and if at any instant of time the electric vector is along the x -axis then the magnetic vector will be along the y -axis.

- (iii) Since $\frac{k}{\omega\mu}$ is a real number, the electric and magnetic vectors are in phase; thus, if at any instant, \mathbf{E} is zero then \mathbf{H} is also zero, similarly when \mathbf{E} attains its maximum value, \mathbf{H} also attains its maximum value, etc.
- (iv) The speed of propagation $v (= \frac{\omega}{k})$ of the wave is given by the following equation.

$$v = \frac{\omega}{k} = \frac{1}{\sqrt{\epsilon\mu}} \quad (35)$$

For free-space

$$\left. \begin{aligned} \epsilon &= \epsilon_0 = 8.8542 \times 10^{-12} \text{ C}^2/\text{N-m}^2 \\ \mu &= \mu_0 = 4\pi \times 10^{-7} \text{ N-sec}^2/\text{C}^2 \end{aligned} \right\} \quad (36)$$

Thus

$$\begin{aligned} c &= \frac{1}{\sqrt{8.8542 \times 10^{-12} \times 4\pi \times 10^{-7}}} \\ &= 2.99794 \times 10^8 \text{ m/sec} \end{aligned} \quad (37)$$

represents the speed of propagation of electromagnetic waves in free space. It may be worthwhile to

mention here that the physical laws described by Eqs (5), (6) and (7) were known before Maxwell; he had only introduced the term $\frac{\partial \mathbf{D}}{\partial t}$ in Eq. (8) (see Sec. 20.9) and it is the presence of this term which leads to the prediction of electromagnetic waves.* Maxwell, using the then known value of ϵ_0 found that the speed of electromagnetic waves should be 3.1074×10^8 m/sec; this he found to be very close to the velocity of light which was, at that time, known to be 3.14858×10^8 m/sec (as measured by Fizeau in 1849). Just based on these two numbers and with 'faith in the rationality of nature' he propounded the electromagnetic theory of light and predicted that light must be an electromagnetic wave. In the words of Maxwell himself, the speed of electromagnetic waves

"... calculated from the electromagnetic measurements of Kohlrausch and Weber, agrees so exactly with the velocity of light calculated from the optical experiments of M. Fizeau, that we can scarcely avoid the inference that *light consists in the transverse undulations of the same medium which is the cause of electric and magnetic phenomena.*"

- (v) The refractive index (n) of a dielectric (characterized by dielectric permittivity ϵ and magnetic permeability μ) would be given by

$$n = \frac{c}{v} = \left(\frac{\epsilon\mu}{\epsilon_0\mu_0} \right)^{1/2} \quad (38)$$

For most dielectrics $\mu \approx \mu_0$ and one obtains

$$n = \sqrt{\epsilon/\epsilon_0} = \sqrt{\kappa} \quad (39)$$

where $\kappa (= \epsilon/\epsilon_0)$ is known as the dielectric constant of the medium.

- (vi) The electric and magnetic waves are interdependent; neither can exist without the other. Physically, an electric field varying in time produces a magnetic field varying in space and time; this changing magnetic field produces an electric field varying in space and time and so on. This mutual generation of electric and magnetic fields result in the propagation of the electromagnetic wave.
- (vii) Since Maxwell's equations are linear in \mathbf{E} and \mathbf{H} , so, if $(\mathbf{E}_1, \mathbf{H}_1)$ and $(\mathbf{E}_2, \mathbf{H}_2)$ are two independent solutions of the Maxwell's equations, then $(\mathbf{E}_1 + \mathbf{E}_2, \mathbf{H}_1 + \mathbf{H}_2)$ will also be a solution of the Maxwell's equations. This is the superposition principle according to which the resultant displacement produced by two indepen-

*It can be easily seen that in the absence of the term $\epsilon \frac{\partial \mathbf{E}}{\partial t}$ on the RHS of Eq. (17) the plane waves, as described by Eqs (18) and (19), will not represent a solution of Eqs (14)–(17).

dent disturbance is the vector sum of the displacements produced by the disturbances independently* (see Chapter 11).

- (viii) The plane wave as represented by Eq. (30) is said to be linearly polarized because the electric vector is always along the x -axis and, similarly, the magnetic vector is always along the y -axis (see Fig. 20.2).

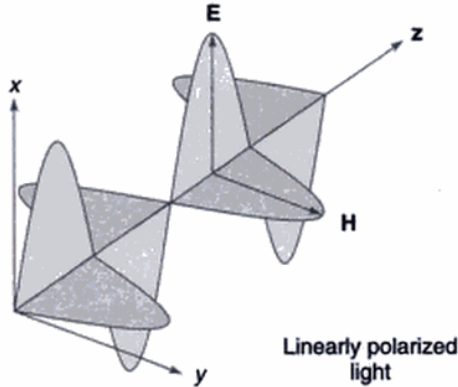


Fig. 20.2 The \mathbf{E} and \mathbf{H} vectors (at a particular instant of time) for a x -polarized wave. The electric vectors always lie in the x - z plane and the magnetic vectors lie in the y - z plane.

However, we may consider two independent plane waves**

$$\left. \begin{aligned} \mathbf{E}_1 &= \hat{\mathbf{x}} E_0 \cos(kz - \omega t) \\ \mathbf{H}_1 &= \hat{\mathbf{y}} H_0 \cos(kz - \omega t) \end{aligned} \right\} \quad (40)$$

and

$$\left. \begin{aligned} \mathbf{E}_2 &= \hat{\mathbf{y}} E_0 \cos\left(kz - \omega t - \frac{\pi}{2}\right) \\ &= \hat{\mathbf{y}} E_0 \sin(kz - \omega t) \\ \mathbf{H}_2 &= -\hat{\mathbf{x}} H_0 \cos\left(kz - \omega t - \frac{\pi}{2}\right) \\ &= -\hat{\mathbf{x}} H_0 \sin(kz - \omega t) \end{aligned} \right\} \quad (41)$$

then

$$\left. \begin{aligned} \mathbf{E} &= \mathbf{E}_1 + \mathbf{E}_2 \\ &= E_0 [\hat{\mathbf{x}} \cos(kz - \omega t) + \hat{\mathbf{y}} \sin(kz - \omega t)] \\ \mathbf{H} &= \mathbf{H}_1 + \mathbf{H}_2 \\ &= H_0 [\hat{\mathbf{y}} \cos(kz - \omega t) - \hat{\mathbf{x}} \sin(kz - \omega t)] \end{aligned} \right\} \quad (42)$$

*Thus the superposition principle is a consequence of the linearity of the Maxwell's equations. If, for example, the fields associated with the electromagnetic wave are so high that the dielectric permittivity ϵ depends on \mathbf{E} itself, then Maxwell's equations will become non-linear and superposition principle will not remain valid. Indeed when we discuss any nonlinear phenomenon, the superposition principle does not hold.

**We are considering the real part of the exponentials appearing on the RHS of Eq. (30).

***See also Sec. 19.4.

will also be a solution which will represent a left circularly polarized wave.***

- (ix) There exists a wide and continuous variation in the frequency (and wavelength) of electromagnetic waves. The electromagnetic spectrum is shown in Fig. 20.3. The radio waves correspond to wavelengths in the range $10 - 1000$ m whereas the wavelengths of X-rays are in the region of Angstroms ($1 \text{ \AA} = 10^{-10} \text{ m}$). The range of wavelengths of various kinds of electromagnetic waves are shown in Fig. 20.3 and as can be seen, the visible region ($4 \times 10^{-5} \text{ cm} < \lambda < 7 \times 10^{-5} \text{ cm}$) occupies a very small portion of the spectrum. The methods of production of different kinds of waves are different; for example, X-rays are usually produced by the sudden stopping or deflection of electrons whereas radio waves may be produced by varying the charge on an antenna. However, all wavelengths propagate with the same speed in vacuum and are always produced by accelerated charges.

20.3 THE THREE-DIMENSIONAL WAVE EQUATION IN A DIELECTRIC

Once again, we consider a non-charged current-free dielectric ($\rho = 0, \mathbf{J} = 0$) and for such a case, using the 'constitutive relations', we had obtained Eqs (14)–(17). Using vector notation, these equations can be written in the following compact form:

$$\text{div } \mathbf{E} = 0 \quad (43)$$

$$\text{div } \mathbf{H} = 0 \quad (44)$$

$$\text{curl } \mathbf{E} = -\mu \frac{\partial \mathbf{H}}{\partial t} \quad (45)$$

$$\text{curl } \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t} \quad (46)$$

In Sec. 20.2 we had shown that plane-wave solutions indeed satisfy the above equations. In this section we will show that the wave equation can be derived from the above equations.

If we take the curl of Eq. (45), we would obtain

$$\text{curl curl } \mathbf{E} = -\mu \frac{\partial}{\partial t} \text{curl } \mathbf{H} = -\epsilon \mu \frac{\partial^2 \mathbf{E}}{\partial t^2} \quad (47)$$

We must choose the positive sign, the negative sign, would make α complex. Thus

$$\text{and } \left. \begin{aligned} \alpha &= \omega \sqrt{\epsilon \mu} \left[\frac{1}{2} + \frac{1}{2} \left(1 + \frac{\sigma^2}{\omega^2 \epsilon^2} \right)^{1/2} \right]^{1/2} \\ \beta &= \frac{\omega \mu \sigma}{2\alpha} \end{aligned} \right\} \quad (101)$$

Now, when k is complex, Eq. (95) becomes

$$\mathbf{E} = \mathbf{E}_0 \exp(-\beta z) \exp[i(\alpha z - \omega t)] \quad (102)$$

which represents an attenuated wave. The attenuation is due to the Joule-loss. For a good conductor*

$$\frac{\sigma}{\omega \epsilon} \gg 1 \quad (103)$$

and one obtains

$$\alpha \approx \beta \approx \left(\frac{\omega \mu \sigma}{2} \right)^{1/2} \quad (104)$$

Indeed if $\frac{\sigma}{\omega \epsilon} \ll 1$ (say ≤ 0.01), the medium can be classified as a dielectric and if $\frac{\sigma}{\omega \epsilon} \gg 1$ (say ≥ 100), the medium can be classified as a conductor. For

$$0.01 \leq \frac{\sigma}{\omega \epsilon} \leq 100,$$

the medium is said to be a quasi-conductor. Thus, depending on the frequency, a particular material can behave as a dielectric or as a conductor. For example, for fresh water $\frac{\epsilon}{\epsilon_0} \approx 80$ and $\sigma = 10^{-3}$ mhos/m. (Both ϵ and σ can be assumed to be constants at low frequencies.) Thus

$$\frac{\sigma}{\epsilon} \approx \frac{10^{-3}}{80 \times 8.85 \times 10^{-12}} \approx 1.4 \times 10^6 \text{ sec}^{-1}$$

For $\omega = 2\pi \times 10 \text{ sec}^{-1}$

$$\frac{\sigma}{\omega \epsilon} \approx 2 \times 10^4$$

and for $\omega = 2\pi \times 10^{10} \text{ sec}^{-1}$

$$\frac{\sigma}{\omega \epsilon} \approx 2 \times 10^{-5}$$

Thus, fresh water behaves as a good conductor for $\nu \leq 10^3 \text{ sec}^{-1}$ and as a dielectric for $\nu \geq 10^7 \text{ sec}^{-1}$. On the other hand, for copper one may assume $\epsilon \approx \epsilon_0$ and $\sigma \approx 5.8 \times 10^7$ mhos/m and for $\omega = 2\pi \times 10^{10} \text{ sec}^{-1}$

$$\frac{\sigma}{\omega \epsilon} \approx \frac{5.8 \times 10^7}{2\pi \times 10^{10} \times 8.9 \times 10^{-12}} \approx 10^8$$

Thus even for such frequencies it behaves as an excellent conductor.

From Eq. (102), it can be easily seen that the field decreases by a factor e in traversing a distance

$$\delta = \frac{1}{\beta},$$

which is known as the penetration depth. For copper,

$$\mu \approx \mu_0 = 4\pi \times 10^{-7} \text{ N/amp}^2$$

and

$$\begin{aligned} \delta &\approx \left(\frac{2}{\omega \mu \sigma} \right)^{1/2} \approx \left(\frac{2}{2\pi \times 10^{10} \times 4\pi \times 10^{-7} \times 5.8 \times 10^7} \right)^{1/2} \\ &\approx \frac{0.065}{\sqrt{\nu}} \text{ m} \end{aligned}$$

Thus for $\nu \approx 100 \text{ sec}^{-1}$, $\delta \approx 0.0065 \text{ m} = 0.65 \text{ cm}$ whereas for $\nu \approx 10^8 \text{ sec}^{-1}$, $\delta \approx 6.5 \times 10^{-6} \text{ m}$ showing that the penetration decreases with increase in frequency.

20.8 THE CONTINUITY CONDITIONS

In this section we will derive the continuity conditions for electric and magnetic fields at the interface of two media. Let us first consider the equation

$$\text{div } \mathbf{B} = 0 \quad (105)$$

At the interface of two media, we consider a pill box which encloses an area ΔS of the interface (see Fig. 20.7). Let the height of the pill box be l .

Now if we integrate $\text{div } \mathbf{B}$ over the cylindrical volume then, using Gauss' theorem, we obtain

$$0 = \int \text{div } \mathbf{B} dV = \oint_{s_1} \mathbf{B} \cdot d\mathbf{a} + \oint_{s_2} \mathbf{B} \cdot d\mathbf{a} + \oint_{s_3} \mathbf{B} \cdot d\mathbf{a}$$

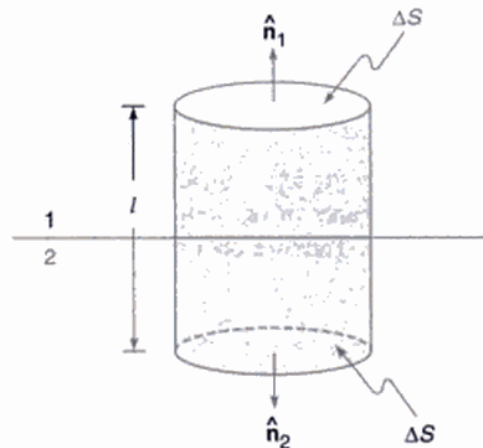


Fig. 20.7 A cylindrical pill box at the interface of two dielectrics.

*The corresponding expressions for an insulator are given in Problem 20.8.

where S_1 and S_2 represent the flat faces of the cylinder and S_3 represents the curved surface of the cylinder. If we let $l \rightarrow 0$, then the third integral vanishes and we obtain

$$\oint_{S_1} \mathbf{B} \cdot d\mathbf{a} = - \oint_{S_2} \mathbf{B} \cdot d\mathbf{a}$$

$$\text{or } \mathbf{B}_1 \cdot \hat{\mathbf{n}}_1 \Delta S = - \mathbf{B}_2 \cdot \hat{\mathbf{n}}_2 \Delta S \quad (106)$$

$$\text{or } B_{1n} = B_{2n} \quad (107)$$

where the directions of $\hat{\mathbf{n}}_1$ and $\hat{\mathbf{n}}_2$ are shown in Fig. 20.7. Thus, the normal component of \mathbf{B} is continuous across the interface.

Similarly, in the absence of free charges

$$\text{div } \mathbf{D} = 0$$

and one obtains*

$$D_{1n} = D_{2n} \quad (108)$$

showing that the normal component of \mathbf{D} is also continuous across the interface.

We next consider the equation

$$\text{curl } \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0$$

We consider a rectangular loop $ABCD$ as shown in Fig. 20.8. Now

$$0 = \oint_S \text{curl } \mathbf{E} \cdot d\mathbf{a} + \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} \quad (109)$$

where the surface integral is over any surface bounding the loop $ABCD$. Using Stokes' theorem, we get

$$\oint \mathbf{E} \cdot d\mathbf{l} = - \int \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} \quad (110)$$

or

$$\left(\int_{AB} + \int_{BC} + \int_{CD} + \int_{DA} \right) \mathbf{E} \cdot d\mathbf{l} = - \int \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a}$$

If we let $l \rightarrow 0$ then the integrals along BC and DA tend to zero and since the area of the loop also tends to zero the RHS also vanishes. Thus we obtain

*Rigorously

$$D_{1n} - D_{2n} = \sigma$$

where σ represents the surface charge density.

**More rigorously,

$$\mathbf{H}_{1t} - \mathbf{H}_{2t}$$

is equal to the normal component of the surface current density. However, if there are no surface currents, which is indeed true for most cases, $H_{1t} = H_{2t}$.

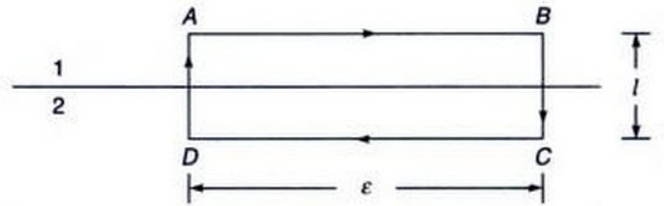


Fig. 20.8 A rectangular loop at the interface of two dielectrics.

$$\int_{AB} \mathbf{E} \cdot d\mathbf{l} + \int_{CD} \mathbf{E} \cdot d\mathbf{l} = 0$$

$$\text{or } (\mathbf{E}_1 \cdot \hat{\mathbf{t}})\epsilon + [\mathbf{E}_2 \cdot (-\hat{\mathbf{t}})]\epsilon = 0$$

$$\text{or } \mathbf{E}_{1t} = \mathbf{E}_{2t}$$

where \mathbf{E}_{1t} and \mathbf{E}_{2t} represent the tangential components of \mathbf{E} which are continuous across the interface.

Similarly, Eq. (8) gives us**

$$\mathbf{H}_{1t} = \mathbf{H}_{2t}$$

In summary, in the absence of any surface current and surface charges, the normal components of \mathbf{B} and \mathbf{D} and the tangential components of \mathbf{H} and \mathbf{E} are continuous across an interface.

20.9 PHYSICAL SIGNIFICANCE OF MAXWELL'S EQUATIONS

Let us first consider the equation

$$\text{div } \mathbf{D} = \rho \quad (111)$$

In free-space

$$\mathbf{D} = \epsilon_0 \mathbf{E} \quad (112)$$

and Eq. (111) becomes

$$\text{div } \mathbf{E} = \frac{\rho}{\epsilon_0} \quad (113)$$

If we integrate the above equation over a volume V , we obtain

$$\int \text{div } \mathbf{E} dV = \frac{1}{\epsilon_0} \int \rho dV$$

Applying the divergence theorem we get

$$\oint \mathbf{E} \cdot d\mathbf{a} = \frac{1}{\epsilon_0} Q \quad (114)$$

which is simply the Gauss' law, [†] i.e. the electric flux through a closed surface is the total charge inside the volume divided by ϵ_0 . In a similar manner, the equation

$$\text{div } \mathbf{B} = 0 \quad (115)$$

gives

$$\oint \mathbf{B} \cdot d\mathbf{a} = 0 \quad (116)$$

i.e., the magnetic flux through a closed surface is always zero; this implies the absence of magnetic monopoles.

We next consider the equation

$$\text{curl } \mathbf{E} = - \frac{\partial \mathbf{B}}{\partial t} \quad (117)$$

which associates a space- and time-dependent electric field with a changing magnetic field. Now, Stokes' theorem tells us that

$$\oint_{\Gamma} \mathbf{E} \cdot d\mathbf{l} = \int_S \text{curl } \mathbf{E} \cdot d\mathbf{a} \quad (118)$$

where the LHS represents a line integral over a closed path Γ and the RHS represents a surface integral over any surface bounding the path Γ . Thus

$$\oint_{\Gamma} \mathbf{E} \cdot d\mathbf{l} = \int_S \text{curl } \mathbf{E} \cdot d\mathbf{a} = - \int_S \frac{\partial \mathbf{B}}{\partial t} \cdot d\mathbf{a} \quad (119)$$

or

$$\oint_{\Gamma} \mathbf{E} \cdot d\mathbf{l} = - \frac{d}{dt} \int_S \mathbf{B} \cdot d\mathbf{a} \quad (120)$$

where in the last step we have used the fact that the surface S is fixed.* The LHS of the above equation represents the induced emf in a closed circuit which is equal to the negative of the rate of change of the magnetic flux through the circuit. This is the famous Faraday's law of induction. It is worthwhile to mention that although this law was discovered by Faraday, it was put in the differential form [see Eq. (117)] by Maxwell.

We now come to the last of the Maxwell's equations i.e.†

$$\text{curl } \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \quad (121)$$

However, Ampere's law (which was known before Maxwell), when expressed as a differential equation, was of the form‡

$$\text{curl } \mathbf{H} = \mathbf{J} \quad (122)$$

which implies that a magnetic field is produced only by currents; for example, if we have a long wire carrying a current, we know that it produces a magnetic field. Since the divergence of the curl of any vector is zero, one obtains

$$\text{div } \mathbf{J} = 0 \quad (123)$$

which may be compared with the equation of continuity

$$\text{div } \mathbf{J} + \frac{\partial \rho}{\partial t} = 0 \quad (124)$$

[†]For a dielectric we would get

$$\oint \mathbf{D} \cdot d\mathbf{a} = Q$$

where

$$\mathbf{D} = \epsilon_0 \mathbf{E} + \mathbf{P}$$

\mathbf{P} being the dipole moment per unit volume. For a linear homogeneous medium,

$$\mathbf{P} = \chi \mathbf{E}$$

where χ is known as the susceptibility. Thus

$$\mathbf{D} = \epsilon \mathbf{E}$$

where

$$\epsilon = \epsilon_0 + \chi$$

is known as the dielectric permittivity of the medium.

*Equation (120) is not valid for a moving system (see, for example, Ref. 3, p. 526).

‡ $\mathbf{H} = \frac{\mathbf{B}}{\mu_0} - \mathbf{M}$, where \mathbf{M} is the magnetic moment/unit volume. For a linear material $\mathbf{M} = \chi_m \mathbf{H}$ and therefore $\mathbf{B} = \mu \mathbf{H}$, where $\mu = \mu_0(1 + \chi_m)$.

‡‡Once again, it was Maxwell, who had expressed Ampere's law as a differential equation.

Thus Eq. (122) is valid only when $\frac{\partial \rho}{\partial t} = 0$. Thus, for the Ampere's law to be consistent with the equation of continuity, Maxwell argued that there must be an additional term on the RHS of Eq. (122).*

The introduction of the term $\frac{\partial \mathbf{D}}{\partial t}$ (which is known as the displacement current) revolutionized physics. Physically it implies that not only a current produces a magnetic field but a changing electric field also produces a magnetic field (as it indeed happens during the charging and discharging of a condenser).† It may be mentioned that it is the presence of the term $\frac{\partial \mathbf{D}}{\partial t}$ which leads to the wave equation (see Sec. 20.3) and, therefore, the prediction of electromagnetic waves. One can thus argue on physical grounds that a changing electric field produces a magnetic field which varies in space and time and this changing magnetic field produces an electric field varying in space and time, and so on. This mutual generation of electric and magnetic fields results in the propagation of electromagnetic waves.

SUMMARY

- In a homogeneous dielectric (with dielectric constant ϵ) Maxwell's equations take the form

$$\text{div } \mathbf{E} = 0$$

$$\text{div } \mathbf{H} = 0$$

$$\text{curl } \mathbf{E} = -\mu_0 \frac{\partial \mathbf{H}}{\partial t}$$

$$\text{curl } \mathbf{H} = \epsilon \frac{\partial \mathbf{E}}{\partial t}$$

where we have assumed the medium to be non-magnetic so that $\mu \approx \mu_0 = 4\pi \times 10^{-7} \text{ N } \text{--} \text{A}^2/\text{C}^2$. The third equation is the Faraday's law. The RHS of the fourth equation is known as the displacement current which was introduced by Maxwell; the inclusion of the displacement current term enabled Maxwell to derive the wave equation

$$\nabla^2 \mathbf{E} = \epsilon \mu_0 \frac{\partial^2 \mathbf{E}}{\partial t^2}$$

- In free space $\epsilon = \epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2/\text{N } \text{--} \text{m}^2$ and therefore the velocity of the electromagnetic waves in free space is given by

$$c = \frac{1}{\sqrt{\epsilon_0 \mu_0}} \approx 3 \times 10^8 \text{ m/s}$$

(The exact value is $2.99792458 \times 10^8 \text{ m/s}$). Maxwell found that the velocity of the electromagnetic waves was very close to the measured velocity of light and with 'faith in rationality of nature' he said that *light is an electromagnetic wave*.

- For a x -polarized electromagnetic wave propagating in the $+z$ direction, we may write

$$\mathbf{E} = \hat{\mathbf{x}} E_0 \cos(kz - \omega t)$$

$$\mathbf{H} = \hat{\mathbf{y}} H_0 \cos(kz - \omega t)$$

$$\text{with } H_0 = \frac{k}{\omega \mu_0} E_0; \frac{\omega}{k} = \frac{1}{\sqrt{\epsilon \mu_0}} = v = \frac{c}{n} \text{ and } n = \sqrt{\frac{\epsilon}{\epsilon_0}}$$

represents the refractive index of the dielectric. The corresponding average energy density is given by

$$\langle u \rangle = \frac{1}{2} \epsilon E_0^2 \quad [\text{J/m}^3]$$

and the intensity is given by

$$I = \frac{1}{2} \epsilon v E_0^2 = \frac{1}{2} \epsilon_0 c n E_0^2 \quad [\text{W/m}^2]$$

For $I = 10^{15} \text{ W/m}^2$, $E_0 \approx 0.9 \times 10^9 \text{ V/m}$; such high electric field can cause spark in air.

- The momentum associated with a plane wave is given by

$$\mathbf{p} = \frac{u}{c} \hat{\mathbf{z}}.$$

PROBLEMS

- 20.1** On the surface of the earth we receive about 1.33 kW of energy per square meter from the sun. Calculate the electric field associated with the sunlight (on the surface of the earth) assuming that it is essentially monochromatic with $\lambda = 6000 \text{ \AA}$.

[Ans. $\sim 1000 \text{ V/m}$]

- 20.2** A 100 W sodium lamp ($\lambda \approx 5890 \text{ \AA}$) is assumed to emit waves uniformly in all directions. What is the radiation pressure on a plane mirror at a distance of 10 m from the bulb?

- 20.3** A 1 kW transmitter is emitting electromagnetic waves (of wavelength 40 m) uniformly in all direc-

*Consequently

$$\text{div curl } \mathbf{H} = 0 = \text{div } \mathbf{J} + \frac{\partial}{\partial t} \text{div } \mathbf{D}$$

or

$$0 = \text{div } \mathbf{J} + \frac{\partial \rho}{\partial t}$$

which is the equation of continuity (we have used Eq. (111)).

†For static fields, $\frac{\partial \mathbf{D}}{\partial t} = 0$ and one obtains Ampere's law.

tions. Calculate the electric field at a distance of 1 km from the transmitter.

- 20.4** Ocean water can be assumed to be a non-magnetic dielectric with $\kappa (= \frac{\epsilon}{\epsilon_0}) = 80$ and $\sigma = 4.3$ mhos/m. (a) Calculate the frequency at which the penetration depth will be 10 cm. (b) Show that for frequencies less than 10^8 sec^{-1} , it can be considered as a good conductor.

[Ans. (a) $\sim 6 \times 10^6 \text{ sec}^{-1}$]

- 20.5** For silver one may assume $\mu \approx \mu_0$ and $\sigma \approx 3 \times 10^7$ mhos/m. Calculate the skin depth at 10^8 sec^{-1} .

[Ans. $\approx 9 \times 10^{-4} \text{ cm}$]

- 20.6** Show that for frequencies $\leq 10^8 \text{ s}^{-1}$, a sample of silicon will act like a good conductor. For silicon one may assume $\frac{\epsilon}{\epsilon_0} \approx 12$ and $\sigma \approx 2$ mhos/m. Also calculate the penetration depth for this sample at $\nu = 10^6 \text{ s}^{-1}$.

- 20.7** In a conducting medium show that \mathbf{H} also satisfies an equation similar to Eq. (94).

- 20.8** Using the analysis given in Sec. 20.7 and assuming $\sigma/\omega\epsilon \ll 1$ (which is valid for an insulator) show that

$$\alpha \approx \omega\sqrt{\epsilon\mu} \left[1 + \frac{1}{8} \left(\frac{\sigma}{\omega\epsilon} \right)^2 \right] = \frac{2\pi}{\lambda_0} n \left[1 + \frac{1}{8} \left(\frac{\sigma}{\omega\epsilon} \right)^2 \right]$$

and

$$\beta \approx \omega\sqrt{\epsilon\mu} \left[\frac{1}{2} \left(\frac{\sigma}{\omega\epsilon} \right) \right] = \frac{2\pi}{\lambda_0} n \left[\frac{1}{2} \left(\frac{\sigma}{\omega\epsilon} \right) \right]$$

where

$$n = \sqrt{\epsilon/\epsilon_0}.$$

- 20.9** For the glass used in a typical optical fiber at $\lambda_0 \approx 8500 \text{ \AA}$, $n = (\epsilon/\epsilon_0)^{1/2} = 1.46$, $\sigma \approx 3.4 \times 10^{-6}$ mhos/m. Calculate $\sigma/\omega\epsilon$ and show that we can use the formulae given in the previous problem. Calculate β and loss in dB/km. [Hint: The power would decrease as $\exp(-2\beta z)$; loss in dB/km is defined in Sec. 24.8.]

[Ans. $\sigma/\omega\epsilon \approx 8 \times 10^{-11}$;
 $\beta \approx 4.3 \times 10^{-4} \text{ m}^{-1}$; loss $\approx 3.7 \text{ dB/km}$]

REFERENCES AND SUGGESTED READING

See at the end of Chapter 21.

Chapter 21

Reflection and Refraction of Electromagnetic Waves

All of electromagnetism is contained in Maxwell's equations... Untold number of experiments have confirmed Maxwell's equations. If we take away the scaffolding be used to build it, we find that Maxwell's beautiful edifice stands on its own.

—Richard Feynman

21.1 INTRODUCTION

In the previous chapter we had discussed the Maxwell's equations and had shown the existence of electromagnetic waves. We had also shown that at an interface, the tangential components of \mathbf{E} and \mathbf{H} and the normal components of \mathbf{D} and \mathbf{B} must be continuous. Using these continuity conditions we will, in this chapter, study the reflection and refraction of plane waves at an interface of two dielectrics (Sec. 21.2) and at an interface of a dielectric and a metal (Sec. 21.3). In Sec. 21.4 we will consider reflectivity (and transmittivity) of a dielectric film.

21.2 REFLECTION AT AN INTERFACE OF TWO DIELECTRICS

Let us consider the incidence of a plane polarized electromagnetic wave on an interface of two media; we assume the plane $x = 0$ to represent the interface. Let (ϵ_1, μ_1) and (ϵ_2, μ_2) represent the dielectric permittivity and magnetic permeability of the media below and above the plane $x = 0$; we will assume both media to be lossless dielectrics, the case of reflection by a conducting surface will be discussed in Sec. 21.3. Let \mathbf{E}_1 , \mathbf{E}_2 and \mathbf{E}_3 denote the electric fields associated with the incident wave, refracted wave and reflected wave respectively. For an incident plane wave, these fields will be of the form

$$\left. \begin{aligned} \mathbf{E}_1 &= \mathbf{E}_{10} \exp[i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)] \\ \mathbf{E}_2 &= \mathbf{E}_{20} \exp[i(\mathbf{k}_2 \cdot \mathbf{r} - \omega_2 t)] \\ \mathbf{E}_3 &= \mathbf{E}_{30} \exp[i(\mathbf{k}_3 \cdot \mathbf{r} - \omega_3 t)] \end{aligned} \right\} \quad (1)$$

where \mathbf{E}_{10} , \mathbf{E}_{20} and \mathbf{E}_{30} are independent of space and time but may, in general, be complex. The vectors \mathbf{k}_1 , \mathbf{k}_2 and \mathbf{k}_3 represent the propagation vectors associated with the

incident, refracted and reflected waves respectively. Since the fields must satisfy Maxwell's equations we must have (see Sec. 20.2):

$$\left. \begin{aligned} \mathbf{k}_1^2 &= \omega^2 \epsilon_1 \mu_1 \\ \mathbf{k}_2^2 &= \omega_2^2 \epsilon_2 \mu_2 \\ \mathbf{k}_3^2 &= \omega_3^2 \epsilon_1 \mu_1 \end{aligned} \right\} \quad (2)$$

As discussed in Sec. 20.9, the fields have to satisfy certain boundary conditions at the interface (which corresponds to $x = 0$) where Eq. (1) takes the form

$$\left. \begin{aligned} \mathbf{E}_1 &= \mathbf{E}_{10} \exp[i(k_{1y} y + k_{1z} z - \omega t)] \\ \mathbf{E}_2 &= \mathbf{E}_{20} \exp[i(k_{2y} y + k_{2z} z - \omega_2 t)] \\ \mathbf{E}_3 &= \mathbf{E}_{30} \exp[i(k_{3y} y + k_{3z} z - \omega_3 t)] \end{aligned} \right\}$$

where k_{1x} , k_{1y} and k_{1z} represent the x , y , and z -components of \mathbf{k}_1 ; similarly for \mathbf{k}_2 and \mathbf{k}_3 . Now, for example, the z -component of the electric field (which is a tangential component) must be continuous at $x = 0$ for *all* values of y , z and t . Consequently, the coefficients of y , z and t in the exponents appearing the above equation must be equal.

Thus

$$\omega = \omega_2 = \omega_3 \quad (3)$$

showing that all the waves have the same frequency. Hence Eqs (2) simplify to

$$k_1^2 = \omega^2 \epsilon_1 \mu_1 = k_3^2 \quad (4)$$

$$k_2^2 = \omega^2 \epsilon_2 \mu_2 \quad (5)$$

Further, we must have

$$k_{1y} = k_{2y} = k_{3y} \quad (6)$$

and

$$k_{1z} = k_{2z} = k_{3z} \quad (7)$$

Without any loss of generality we may choose the y -axis such that

$$k_{1y} = 0$$

(i.e., \mathbf{k}_1 is assumed to lie in the x - z plane—see Fig. 21.1). Consequently,

$$k_{2y} = k_{3y} = 0 \quad (8)$$

Eq. (8) implies that the vectors \mathbf{k}_1 , \mathbf{k}_2 and \mathbf{k}_3 will lie in the same plane. Further, from Eq. (7), we get

$$k_1 \sin \theta_1 = k_2 \sin \theta_2 = k_3 \sin \theta_3 \quad (9)$$

Since $k_1 = k_3$ (see Eq. (4)) we must have $\theta_1 = \theta_3$, i.e., the angle of incidence is equal to angle of reflection. Further,

$$\frac{\sin \theta_1}{\sin \theta_2} = \left(\frac{\epsilon_2 \mu_2}{\epsilon_1 \mu_1} \right)^{1/2} \quad (10)$$

If $v_1 \left(= \frac{1}{\sqrt{\epsilon_1 \mu_1}} \right)$ and $v_2 \left(= \frac{1}{\sqrt{\epsilon_2 \mu_2}} \right)$ represent the speeds of propagation of the waves in media 1 and 2, then*

$$\frac{\sin \theta_1}{\sin \theta_2} = \frac{v_1}{v_2} = \frac{n_2}{n_1} \quad (11)$$

where $n_1 \left(= \frac{c}{v_1} = c \sqrt{\epsilon_1 \mu_1} \right)$

and $n_2 \left(= \frac{c}{v_2} = c \sqrt{\epsilon_2 \mu_2} \right)$ (12)

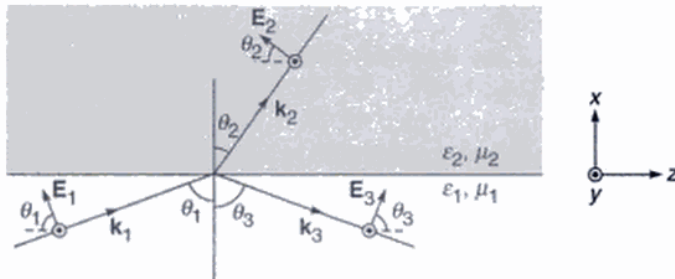


Fig. 21.1 The reflection of a plane wave with its electric vector parallel to the plane of incidence.

*Equation (11) remains valid even when the second medium is anisotropic. As a simple example, if we assume the second medium to be uniaxial with its optic axis along the normal then for the extraordinary wave we would have

$$n_1 \sin \theta_1 = n_{we}(\theta_2) \sin \theta_2$$

where n_{we} would be given by Eq. (96) of Chapter 19 with ψ replaced by θ_2 ; θ_2 represents the direction of \mathbf{k}_2 not of the ray. The above equation would determine θ_2 . (see, e.g., Chapter 3 of Ref. 5).

**We will show later that if the electric vector associated with the incident wave lies in the plane of incidence then the electric vectors associated with the reflected and transmitted waves also lie in the plane of incidence. Similarly, if the electric vector associated with the incident wave is normal to the plane of incidence then the electric vectors associated with the reflected and transmitted waves also lie normal to the plane of incidence—see the discussion just before Example 21.5.

represent the refractive indices of media 1 and 2 respectively. Equation (12) is the well-known Snell's law.

We will now derive expressions for the reflection and transmission coefficients when a plane polarized wave is incident on an interface of two dielectrics. We will first consider the case when the electric vector lies in the plane of incidence which will be followed by the case when the electric vector is at right angles to the plane of incidence.

Case 1. E Parallel to the Plane of Incidence

We will assume the electric vector to lie in the plane of incidence** as shown in Fig. 21.1. The magnetic vectors are along the y -axis. Clearly, the z -component of the electric field represents a tangential component which should be continuous across the surface. Thus

$$E_{1z} + E_{3z} = E_{2z}$$

$$\text{or} \quad -E_1 \cos \theta_1 + E_3 \cos \theta_1 = -E_2 \cos \theta_2 \quad (13)$$

$$\begin{aligned} & [-E_{10} \exp \{i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)\} \\ & + E_{30} \exp \{i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)\}]_{x=0} \cos \theta_1 \\ & = [-E_{20} \exp \{i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)\}]_{x=0} \cos \theta_2 \end{aligned} \quad (14)$$

Once again, since this condition has to be satisfied at all space points in the plane $x = 0$ and at all times, the exponents must be identically equal which leads to Eqs (3), (6) and (7). Thus

$$[E_{10} - E_{30}] \cos \theta_1 = E_{20} \cos \theta_2 \quad (15)$$

Further, the normal component of \mathbf{D} must also be continuous and since $\mathbf{D} = \epsilon \mathbf{E}$, we must have

$$\epsilon_1 E_{1x} + \epsilon_1 E_{3x} = \epsilon_2 E_{2x}$$

or

$$\epsilon_1 [E_{10} + E_{30}] \sin \theta_1 = \epsilon_2 E_{20} \sin \theta_2 \quad (16)$$

Substituting for E_{20} from Eq. (15), we get

$$\epsilon_1 [E_{10} + E_{30}] \sin \theta_1 = \epsilon_2 \sin \theta_2 \frac{[E_{10} - E_{30}]}{\cos \theta_2} \cos \theta_1$$

or

$$[\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2] E_{30} \\ = [\epsilon_2 \sin \theta_2 \cos \theta_1 - \epsilon_1 \sin \theta_1 \cos \theta_2] E_{10}$$

Thus $r_{||} = \frac{E_{30}}{E_{10}}$

$$= \frac{\epsilon_2 \sin \theta_2 \cos \theta_1 - \epsilon_1 \sin \theta_1 \cos \theta_2}{\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2} \quad (17)$$

where $r_{||}$ denotes the amplitude reflection coefficient, the subscript || refers to the fact that we are referring to parallel polarization. If we now divide Eq. (15) by E_{10} and substitute the expression for $\frac{E_{30}}{E_{10}}$ from Eq. (17), we would get

$$\left[1 - \frac{\epsilon_2 \sin \theta_2 \cos \theta_1 - \epsilon_1 \sin \theta_1 \cos \theta_2}{\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2} \right] \cos \theta_1 \\ = \frac{E_{20}}{E_{10}} \cos \theta_2$$

or

$$t_{||} = \frac{E_{20}}{E_{10}} \\ = \frac{2 \epsilon_1 \sin \theta_1 \cos \theta_1}{\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2} \quad (18)$$

where $t_{||}$ denotes the amplitude transmission coefficient.

In order to calculate the reflection coefficient we must determine the ratio of the x -components of the Poynting vectors (See Sec. 20.4) associated with the reflected and transmitted waves. The reason why we should take the ratio of the x -component can easily be understood by referring to Fig. 21.2. If S_1 denotes the magnitude of the Poynting vector associated with the incident wave then the energy incident on the area dA (on the surface $x = 0$) per unit time would be $S_{1x} dA = S_1 dA \cos \theta_1$. Similarly, the energy transmitted through the area dA would be

$$S_{2x} dA = S_2 \cos \theta_2 dA$$

and the energy reflected from the area dA would be

$$S_{3x} dA = S_3 \cos \theta_1 dA$$

If $R_{||}$ and $T_{||}$ denote the reflection and transmission coefficients, then*

$$R_{||} = \frac{S_{3x}}{S_{1x}} = \frac{S_3 \cos \theta_1}{S_1 \cos \theta_1} \quad (19) \\ = \frac{\langle \mathbf{E}_3 \times \mathbf{H}_3 \rangle}{\langle \mathbf{E}_1 \times \mathbf{H}_1 \rangle} = \frac{\sqrt{\epsilon_1/\mu_1} |E_{30}|^2}{\sqrt{\epsilon_1/\mu_1} |E_{10}|^2} \quad [\text{see Sec. 20.4}]$$

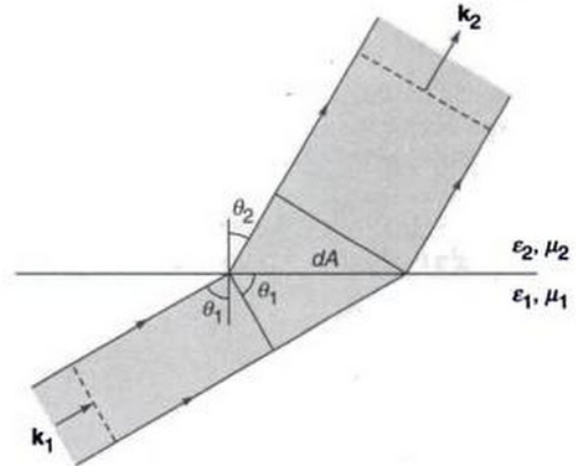


Fig. 21.2 If the cross-sectional area of the incident beam is $dA \cos \theta_1$ then the cross-sectional area of the transmitted beam is $dA \cos \theta_2$ where θ_1 and θ_2 represent the angles of incidence and refraction respectively.

$$= \left| \frac{E_{30}}{E_{10}} \right|^2$$

or

$$R_{||} = \left[\frac{\epsilon_2 \sin \theta_2 \cos \theta_1 - \epsilon_1 \sin \theta_1 \cos \theta_2}{\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2} \right]^2 \quad (20)$$

and

$$T_{||} = \frac{S_{2x}}{S_{1x}} = \frac{S_2 \cos \theta_2}{S_1 \cos \theta_1} \\ = \frac{\langle \mathbf{E}_2 \times \mathbf{H}_2 \rangle \cos \theta_2}{\langle \mathbf{E}_1 \times \mathbf{H}_1 \rangle \cos \theta_1} \\ = \frac{\sqrt{\epsilon_2/\mu_2} |E_{20}|^2 \cos \theta_2}{\sqrt{\epsilon_1/\mu_1} |E_{10}|^2 \cos \theta_1} \\ = \sqrt{\epsilon_2/\epsilon_1} \sqrt{\mu_1/\mu_2} \frac{\sin \theta_2}{\sin \theta_1}$$

$$\left[\frac{2 \epsilon_1 \sin \theta_1 \cos \theta_1}{\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2} \right]^2 \frac{\cos \theta_2}{\cos \theta_1}$$

where we have substituted for $\sqrt{\mu_1/\mu_2}$ from Eqs (11) and (12). Thus

$$T_{||} = \frac{4 \epsilon_1 \epsilon_2 \sin \theta_1 \sin \theta_2 \cos \theta_1 \cos \theta_2}{[\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2]^2} \quad (21)$$

*To calculate the Poynting vector we must use the real parts of \mathbf{E} and \mathbf{H} ; see Sec. 20.4

It can easily be seen that

$$R_{\parallel} + T_{\parallel} = 1 \quad (22)$$

For non-magnetic media, $\mu_1 \approx \mu_2 \approx \mu_0 = 4\pi \times 10^{-7} \text{ N/amp}^2$ and the expression for the amplitude reflection coefficient [Eq. (17)] simplifies to*

$$r_{\parallel} = \frac{n_2^2 \sin \theta_2 \cos \theta_1 - n_1^2 \sin \theta_1 \cos \theta_2}{n_2^2 \sin \theta_2 \cos \theta_1 + n_1^2 \sin \theta_1 \cos \theta_2} \quad (23)$$

Since $n_1 \sin \theta_1 = n_2 \sin \theta_2$

we get

$$r_{\parallel} = \frac{n_2 \cos \theta_1 - n_1 \cos \theta_2}{n_2 \cos \theta_1 + n_1 \cos \theta_2} \quad (24a)$$

$$= \frac{\sin \theta_1 \cos \theta_1 - \sin \theta_2 \cos \theta_2}{\sin \theta_1 \cos \theta_1 + \sin \theta_2 \cos \theta_2} \quad (24b)$$

or

$$\begin{aligned} r_{\parallel} &= \frac{\sin 2\theta_1 - \sin 2\theta_2}{\sin 2\theta_1 + \sin 2\theta_2} \\ &= \frac{2 \cos(\theta_1 + \theta_2) \sin(\theta_1 - \theta_2)}{2 \sin(\theta_1 + \theta_2) \cos(\theta_1 - \theta_2)} \\ &= \frac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)} \end{aligned} \quad (24c)$$

Similarly, starting from Eq. (18) one easily obtains

$$t_{\parallel} = \frac{2 \cos \theta_1 \sin \theta_2}{\sin(\theta_1 + \theta_2) \cos(\theta_1 - \theta_2)} \quad (25)$$

From Eqs (24) and (25) we may deduce the following:

(a) No Reflection when $n_2 = n_1$

When $n_2 = n_1$, $\theta_2 = \theta_1$ and we get

$$r_{\parallel} = 0 \text{ and } t_{\parallel} = 1$$

Thus there is no reflection when the second medium has the same refractive index as the first medium (obviously!). Thus if we have a transparent solid immersed in a liquid of the same refractive index, the solid would not be seen!

*We are using here the fact that for nonmagnetic media

$$n = \frac{c}{v} = \sqrt{\epsilon \mu / \epsilon_0 \mu_0} \approx \sqrt{\epsilon / \epsilon_0}$$

Thus

$$n^2 = \epsilon / \epsilon_0$$

**Consequently the entire energy will appear in the transmitted beam. But $t_{\parallel} = 2 \cos^2 \theta_1$ —why? (See Fig. 21.2).

***In Ref. 4, a beautiful physical argument has been given as to why the reflected light should be linearly polarized when the angle of incidence corresponds to the Brewster angle.

(b) Polarization by Reflection: Brewster's Law

If the angle of incidence is such that

$$\theta_1 + \theta_2 = \frac{\pi}{2}, \text{ then } r_{\parallel} = 0$$

i.e. there is no reflected beam.** Thus, if an unpolarized beam is incident at an angle such that $\theta_1 + \theta_2 = \frac{\pi}{2}$, then the parallel component of the E-vector will not be reflected and the reflected light will be polarized with its E-vector perpendicular to the plane of incidence (see Fig. 21.3). This is the famous *Brewster's Law*. The corresponding angle of incidence is known as the Brewster angle (or the polarizing angle) and is usually denoted by θ_p .

Notice that the angle of refraction will be $\frac{\pi}{2} - \theta_p$ and therefore Snell's law takes the form

$$\frac{n_2}{n_1} = \frac{\sin \theta_1}{\sin \theta_2} = \frac{\sin \theta_p}{\sin \left(\frac{\pi}{2} - \theta_p \right)} = \tan \theta_p \quad (26)$$

or

$$\theta_p = \tan^{-1} \left(\frac{n_2}{n_1} \right) \quad (27)$$

Thus, when the angle of incidence is equal to $\tan^{-1} \left(\frac{n_2}{n_1} \right)$ then the reflected beam is plane polarized. Further, the transmitted beam is partially polarized. It is easily seen that at the polarizing angle, the reflected ray is at right angles to the refracted ray.***

(c) Phase Change on Reflection and Stokes' Relations

When light is incident on a denser medium, $\theta_2 < \theta_1$ and for $(\theta_1 + \theta_2) > \frac{\pi}{2}$ (i.e. $\theta_1 > \theta_p$), r_{\parallel} is negative implying a phase change of π . However, no such phase change occurs when $\theta_1 < \theta_p$. We will discuss this point in detail later.

The amplitude reflection and transmission coefficients satisfy Stokes' relations (see Example 21.1).

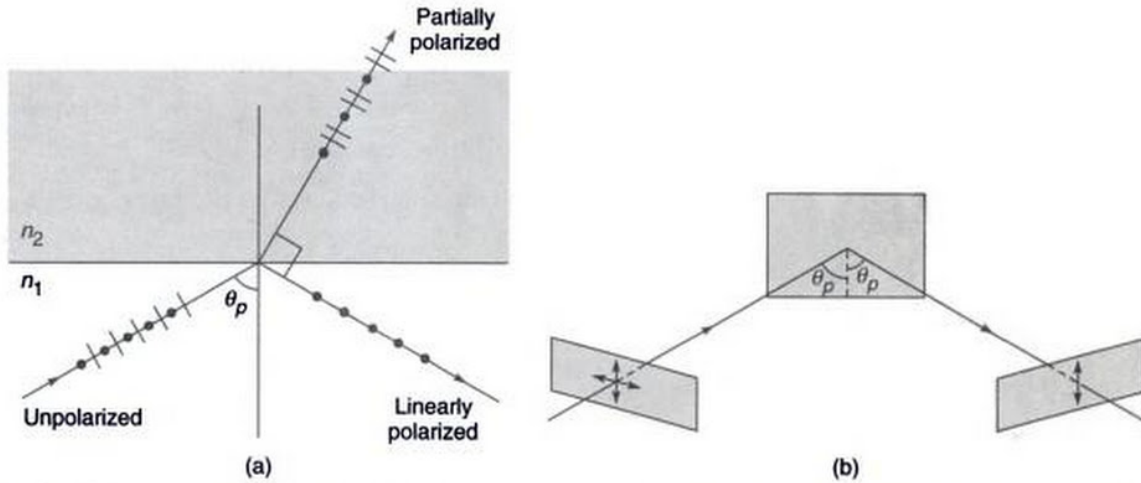


Fig. 21.3 When an unpolarized beam of light is incident on a dielectric at the polarizing angle [i.e., the angle of incidence is equal to $\tan^{-1}(n_2/n_1)$] then the reflected beam is plane-polarized with its E-vector perpendicular to the plane of incidence. The transmitted beam is partially polarized. The dashed line in (b) is normal to the reflecting surface.

(d) Reflection at Grazing Incidence

For grazing incidence ($\theta_1 \approx \frac{\pi}{2}$), Eq. (23) can be written in the form*

$$r_{\parallel} = \frac{\frac{\sin \theta_1}{\sin \theta_2} \sin \alpha_1 - \sin \alpha_2}{\frac{\sin \theta_1}{\sin \theta_2} \sin \alpha_1 + \sin \alpha_2} = \frac{n \sin \alpha_1 - \sin \alpha_2}{n \sin \alpha_1 + \sin \alpha_2} \quad (28)$$

where $n = \frac{n_2}{n_1}$, $\alpha_1 = \frac{\pi}{2} - \theta_1$ and $\alpha_2 = \frac{\pi}{2} - \theta_2$ and at grazing incidence both these angles will be small. Now

$$n = \frac{\sin \theta_1}{\sin \theta_2} = \frac{\cos \alpha_1}{\cos \alpha_2}$$

or

$$\sin \alpha_2 = [1 - \cos^2 \alpha_2]^{1/2} = \left[1 - \frac{\cos^2 \alpha_1}{n^2}\right]^{1/2}$$

Thus

$$\begin{aligned} r_{\parallel} &= \frac{n \sin \alpha_1 - \left[1 - \frac{\cos^2 \alpha_1}{n^2}\right]^{1/2}}{n \sin \alpha_1 + \left[1 - \frac{\cos^2 \alpha_1}{n^2}\right]^{1/2}} \\ &\approx \frac{n \alpha_1 - \left[1 - \frac{1}{n^2}\right]^{1/2}}{n \alpha_1 + \left[1 - \frac{1}{n^2}\right]^{1/2}} \quad (29) \end{aligned}$$

where we have replaced $\sin \alpha_1$ by α_1 and $\cos \alpha_1$ by 1 (thus we have retained terms proportional to α_1 but neglected terms of higher order—this will be justified when α_1 is small). Thus

$$\begin{aligned} r_{\parallel} &= - \left[1 - \frac{n \alpha_1}{\sqrt{(n^2 - 1)/n^2}}\right] \left[1 + \frac{n \alpha_1}{\sqrt{(n^2 - 1)/n^2}}\right]^{-1} \\ &\approx - \left[1 - \frac{2 n^2 \alpha_1}{\sqrt{n^2 - 1}}\right] \rightarrow -1 \text{ as } \alpha_1 \rightarrow 0 \quad (30) \end{aligned}$$

which shows that the reflection is complete at grazing incidence. The transmission coefficient tends to zero as it is indeed obvious from Eq. (25). Thus, if we hold a glass plate horizontally at the level of the eye (see Fig. 21.4) the angle of incidence will be close to $\pi/2$ and the plate will act as a mirror.

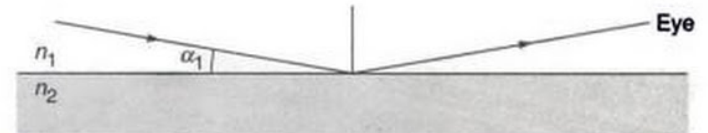


Fig. 21.4 When light is incident at grazing angle (i.e., $\alpha_1 \approx 0$) the reflection is almost complete.

(e) Total Internal Reflection

When an electromagnetic wave is incident on a rarer medium (i.e., $n_2 < n_1$) then $\theta_2 > \theta_1$ and Snell's law [Eq. (12)] can be written in the form

*The second medium must be a denser medium (i.e., $n_2 > n_1$) otherwise the beam will undergo total internal reflection [see part (e)].

$$\begin{aligned}
 1 + r_{||}' &= 1 - r_{||}^2 \\
 &= 1 - \frac{(\epsilon_2 \sin \theta_2 \cos \theta_1 - \epsilon_1 \sin \theta_1 \cos \theta_2)^2}{(\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2)^2} \\
 &= \frac{4\epsilon_1 \epsilon_2 \sin \theta_1 \cos \theta_1 \sin \theta_2 \cos \theta_2}{(\epsilon_2 \sin \theta_2 \cos \theta_1 + \epsilon_1 \sin \theta_1 \cos \theta_2)^2}
 \end{aligned}$$

Now $t_{||}$ is given by Eq. (18); if we make the above mentioned replacements we would get

$$t_{||}' = \frac{2\epsilon_2 \sin \theta_2 \cos \theta_2}{\epsilon_1 \sin \theta_1 \cos \theta_2 + \epsilon_2 \sin \theta_2 \cos \theta_1} \quad (41)$$

If we multiply the above expression for $t_{||}'$ by Eq. (18) we would readily get Eq. (39).

Example 21.2 In deriving the reflection and transmission coefficients, instead of assuming the continuity of the normal component of \mathbf{D} , if we assume the continuity of tangential component of \mathbf{H} show that the same results for the reflection and transmission coefficients would be obtained.

Solution: It is obvious from Fig. 21.1 that the magnetic field will be in the y -direction* which represents a tangential component. Thus if \mathbf{H}_1 , \mathbf{H}_2 and \mathbf{H}_3 represent the magnetic fields associated with the incident, transmitted and reflected waves respectively, then we may write

$$\left. \begin{aligned} \mathbf{H}_1 &= \hat{\mathbf{y}} H_{10} \exp[i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)] \\ \mathbf{H}_2 &= \hat{\mathbf{y}} H_{20} \exp[i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] \\ \mathbf{H}_3 &= \hat{\mathbf{y}} H_{30} \exp[i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)] \end{aligned} \right\} \quad (42)$$

Continuity of the y -component of the field would give

$$H_{10} + H_{30} = H_{20} \quad (43)$$

But

$$\mathbf{H} = \frac{\mathbf{k} \times \mathbf{E}}{\omega \mu} \quad (44)$$

Thus

$$\frac{k_1}{\omega \mu_1} (E_{10} + E_{30}) = \frac{k_2}{\omega \mu_2} E_{20} \quad (45)$$

Continuity of the tangential component of \mathbf{E} gave (see Eq. (15))

$$\begin{aligned}
 (E_{10} - E_{30}) \cos \theta_1 &= E_{20} \cos \theta_2 \\
 &= \frac{k_1}{k_2} \frac{\mu_2}{\mu_1} (E_{10} + E_{30}) \cos \theta_2
 \end{aligned}$$

Thus

$$r_{||} = \frac{E_{30}}{E_{10}} = \frac{\frac{k_2}{\mu_2} \cos \theta_1 - \frac{k_1}{\mu_1} \cos \theta_2}{\frac{k_2}{\mu_2} \cos \theta_1 + \frac{k_1}{\mu_1} \cos \theta_2} \quad (46)$$

*The vector $\mathbf{E} \times \mathbf{H}$ is along the direction of propagation.

$$= \frac{\sqrt{\epsilon_2/\mu_2} \cos \theta_1 - \sqrt{\epsilon_1/\mu_1} \cos \theta_2}{\sqrt{\epsilon_2/\mu_2} \cos \theta_1 + \sqrt{\epsilon_1/\mu_1} \cos \theta_2} \quad (47)$$

If we now use Snell's law, i.e.

$$\frac{\sin \theta_1}{\sin \theta_2} = \left(\frac{\epsilon_2 \mu_2}{\epsilon_1 \mu_1} \right)^{1/2}$$

we would get Eq. (17). However, from Eq. (47) we get the reflection coefficient at normal incidence

$$r_{||} = \frac{\sqrt{\epsilon_2/\mu_2} - \sqrt{\epsilon_1/\mu_1}}{\sqrt{\epsilon_2/\mu_2} + \sqrt{\epsilon_1/\mu_1}} \approx \frac{n_2 - n_1}{n_2 + n_1} \quad (48)$$

the last relation holds only for nonmagnetic media ($\mu_2 \approx \mu_1 \approx \mu_0$). Thus

$$R = \left| \frac{E_{30}}{E_{10}} \right|^2 = \left(\frac{n_2 - n_1}{n_2 + n_1} \right)^2 \quad (49)$$

For a beam incident from air onto glass $n_1 = 1.0$, $n_2 = 1.5$ and therefore,

$$R = 0.04 \quad (50)$$

Thus about 4% of the light is reflected and 96% is transmitted into glass.

Example 21.3 Calculate the phase change in the beam which undergoes total internal reflection.

Solution:

$$r_{||} = \frac{\cos \theta_1 - i\gamma}{\cos \theta_1 + i\gamma} \quad (51)$$

$$= \frac{Ae^{-i\phi}}{Ae^{i\phi}} = e^{-2i\phi}$$

where

$$A = [\cos^2 \theta_1 + \gamma^2]^{1/2}$$

$$\cos \phi = \frac{\cos \theta_1}{[\cos^2 \theta_1 + \gamma^2]^{1/2}}, \quad \sin \phi = \frac{\gamma}{[\cos^2 \theta_1 + \gamma^2]^{1/2}}$$

Thus

$$E_{30} = E_{10} e^{-2i\phi} \quad (52)$$

and the phase change (Δ) is given by

$$\begin{aligned}
 \Delta &= 2\phi = 2 \tan^{-1} \frac{\gamma}{\cos \theta_1} \\
 &= 2 \tan^{-1} \left[\frac{\epsilon_1}{\epsilon_2} \frac{\sqrt{\sin^2 \theta_1 - \sin^2 \theta_c}}{\cos \theta_1} \right] \quad (53)
 \end{aligned}$$

Example 21.4 Determine the nature of the transmitted wave when the beam undergoes total internal reflection.

Solution: The electric field associated with the transmitted wave is given by (see Eq. (1)):

$$\begin{aligned} \mathbf{E}_2 &= \mathbf{E}_{20} \exp [i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] \\ &= \mathbf{E}_{20} \exp [i(k_{2x}x + k_{2z}z - \omega t)] \\ &= \mathbf{E}_{20} \exp [i(k_2 x \cos \theta_2 + k_2 z \sin \theta_2 - \omega t)] \end{aligned} \quad (54)$$

[see Fig. 21.1]. Now

$$\frac{\sin \theta_1}{\sin \theta_2} = \sqrt{\frac{\epsilon_2}{\epsilon_1}}$$

$$\therefore \sin \theta_2 = \sqrt{\frac{\epsilon_1}{\epsilon_2}} \sin \theta_1$$

and

$$\begin{aligned} \cos \theta_2 &= \sqrt{1 - \frac{\epsilon_1}{\epsilon_2} \sin^2 \theta_1} \\ &= \sqrt{\frac{\epsilon_1}{\epsilon_2}} \sqrt{\frac{\epsilon_2}{\epsilon_1} - \sin^2 \theta_1} \\ &= \sqrt{\frac{\epsilon_1}{\epsilon_2}} i\gamma \end{aligned}$$

Thus

$$\mathbf{E}_2 = \mathbf{E}_{20} e^{-\beta x} \exp \left[i \left\{ \left(k_2 \sqrt{\frac{\epsilon_1}{\epsilon_2}} \sin \theta_1 \right) z - \omega t \right\} \right] \quad (55)$$

where

$$\beta = k_2 \sqrt{\frac{\epsilon_1}{\epsilon_2}} \gamma = \frac{\omega}{c} \sqrt{n_1^2 \sin^2 \theta_1 - n_2^2} \quad (56)$$

The field given by Eq. (55) represents a wave propagating in the $+z$ -direction with an amplitude decreasing exponentially in the x -direction. Such a wave is known as a 'surface-wave' or an 'evanescent wave' (see Fig. 21.6). Such waves have many interesting applications.*

Case 2. \mathbf{E} Perpendicular to the Plane of Incidence

Let us next consider the reflection and refraction of a linearly polarized plane wave with its electric vector perpendicular

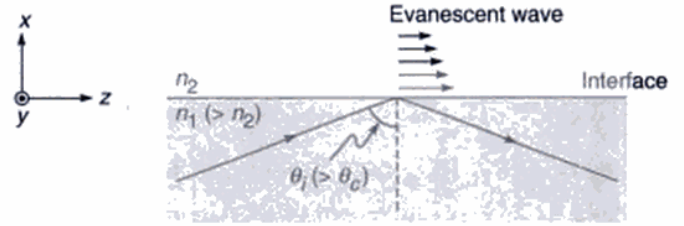


Fig. 21.6 An evanescent wave is generated in the rarer medium when a beam undergoes total internal reflection. The evanescent wave propagates along the z -axis and the amplitude decreases along the x -axis.

to the plane of incidence; the reflection is assumed to occur at the interface of two dielectrics. Thus the electric vectors will be along the y -axis (see Fig. 21.7) and we may write

$$\left. \begin{aligned} \mathbf{E}_1 &= \hat{\mathbf{y}} E_{10} \exp [i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)] \\ \mathbf{E}_2 &= \hat{\mathbf{y}} E_{20} \exp [i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] \\ \mathbf{E}_3 &= \hat{\mathbf{y}} E_{30} \exp [i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)] \end{aligned} \right\} \quad (57)$$

where \mathbf{E}_1 , \mathbf{E}_2 and \mathbf{E}_3 denote the electric vectors associated with the incident, transmitted and reflected waves respectively. Since the y -axis is tangential to the interface, the y -component of \mathbf{E} must be continuous across the interface; consequently

$$E_{10} + E_{30} = E_{20} \quad (58)$$

The directions of the magnetic fields** are also shown in Fig. 21.7; they lie in the plane of incidence and are given by

$$\left. \begin{aligned} \mathbf{H}_1 &= \mathbf{H}_{10} \exp [i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)] \\ &= \frac{\mathbf{k}_1 \times \mathbf{E}_{10}}{\omega \mu_1} \exp [i(\mathbf{k}_1 \cdot \mathbf{r} - \omega t)] \\ \mathbf{H}_2 &= \mathbf{H}_{20} \exp [i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] \\ &= \frac{\mathbf{k}_2 \times \mathbf{E}_{20}}{\omega \mu_2} \exp [i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] \\ \mathbf{H}_3 &= \mathbf{H}_{30} \exp [i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)] \\ &= \frac{\mathbf{k}_3 \times \mathbf{E}_{30}}{\omega \mu_1} \exp [i(\mathbf{k}_3 \cdot \mathbf{r} - \omega t)] \end{aligned} \right\} \quad (59)$$

(Notice that \mathbf{H} lies in the plane of incidence.) Since \mathbf{k}_1 is at right angles to \mathbf{E}_{10} , the magnitude of \mathbf{H}_{10} is simply $\frac{k_1 E_{10}}{\omega \mu_1}$; similarly for \mathbf{H}_{20} and \mathbf{H}_{30} . It is obvious from Fig. 21.7 that

*See, for example, Ref. 2.

**It may be noted that since the displacement vector \mathbf{D} has no component normal to the interface the continuity of normal component of \mathbf{D} will not give us any equation.

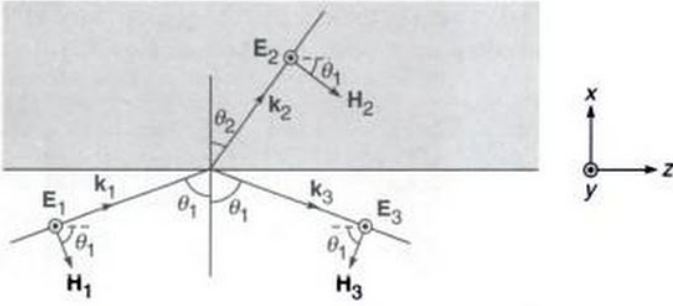


Fig. 21.7 The reflection and refraction of a plane wave with the electric vector lying perpendicular to the plane of incidence.

for the z -component of the magnetic field to be continuous, we must have

$$H_{10} \cos \theta_1 - H_{30} \cos \theta_1 = H_{20} \cos \theta_2 \quad (60)$$

or

$$\frac{k_1}{\omega \mu_1} (E_{10} - E_{30}) \cos \theta_1 = \frac{k_2}{\omega \mu_2} E_{20} \cos \theta_2 \quad (61)$$

Substituting the expression for E_{20} from Eq. (58) we get

$$\frac{k_1}{\omega \mu_1} (E_{10} - E_{30}) \cos \theta_1 = \frac{k_2}{\omega \mu_2} (E_{10} + E_{30}) \cos \theta_2$$

Rearranging, we get

$$r_{\perp} = \frac{E_{30}}{E_{10}} = \frac{\frac{k_1}{\omega \mu_1} \cos \theta_1 - \frac{k_2}{\omega \mu_2} \cos \theta_2}{\frac{k_1}{\omega \mu_1} \cos \theta_1 + \frac{k_2}{\omega \mu_2} \cos \theta_2} \quad (62)$$

$$= \frac{\sqrt{\epsilon_1/\mu_1} \cos \theta_1 - \sqrt{\epsilon_2/\mu_2} \cos \theta_2}{\sqrt{\epsilon_1/\mu_1} \cos \theta_1 + \sqrt{\epsilon_2/\mu_2} \cos \theta_2} \quad (63)$$

$$\approx \frac{\sin \theta_2 \cos \theta_1 - \sin \theta_1 \cos \theta_2}{\sin \theta_2 \cos \theta_1 + \sin \theta_1 \cos \theta_2} = -\frac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)} \quad (64)$$

Further

$$t_{\perp} = \frac{E_{20}}{E_{10}} = 1 + \frac{E_{30}}{E_{10}} = \frac{2\sqrt{\epsilon_1/\mu_1} \cos \theta_1}{\sqrt{\epsilon_1/\mu_1} \cos \theta_1 + \sqrt{\epsilon_2/\mu_2} \cos \theta_2} \quad (65)$$

$$\approx \frac{2 \sin \theta_2 \cos \theta_1}{\sin(\theta_1 + \theta_2)} \quad (66)$$

where the subscript \perp on r and t refers to the fact that we are referring to the state of polarization in which the \mathbf{E} -vector is perpendicular to the plane of incidence. It may be mentioned that Eqs (62), (63) and (65) are exact, whereas Eqs (64) and (66) are valid for non-magnetic media. Once again, one can show that when $\theta_1 > \theta_c$, total internal reflection will occur and for grazing incidence the reflection is complete.

We may summarize now the amplitude reflection and transmission coefficients for the two cases; the results being valid for non-magnetic media:

$$r_{\parallel} = \frac{n_2 \cos \theta_1 - n_1 \cos \theta_2}{n_2 \cos \theta_1 + n_1 \cos \theta_2} = \frac{(n_2/n_1)^2 \cos \theta_1 - \sqrt{(n_2/n_1)^2 - \sin^2 \theta_1}}{(n_2/n_1)^2 \cos \theta_1 + \sqrt{(n_2/n_1)^2 - \sin^2 \theta_1}} \quad (67)$$

$$= \frac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)} \quad (68)$$

$$r_{\perp} = \frac{n_1 \cos \theta_1 - n_2 \cos \theta_2}{n_1 \cos \theta_1 + n_2 \cos \theta_2} = \frac{\cos \theta_1 - \sqrt{(n_2/n_1)^2 - \sin^2 \theta_1}}{\cos \theta_1 + \sqrt{(n_2/n_1)^2 - \sin^2 \theta_1}} \quad (69)$$

$$= -\frac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)} \quad (70)$$

$$t_{\parallel} = \frac{2n_1 \cos \theta_1}{n_2 \cos \theta_1 + n_1 \cos \theta_2} = \frac{2 \cos \theta_1 \sin \theta_2}{\sin \theta_1 \cos \theta_1 + \sin \theta_2 \cos \theta_2} \quad (71)$$

$$t_{\perp} = \frac{2n_1 \cos \theta_1}{n_1 \cos \theta_1 + n_2 \cos \theta_2} = \frac{2 \cos \theta_1 \sin \theta_2}{\sin(\theta_1 + \theta_2)} \quad (72)$$

Eqs (67)–(72) are known as Fresnel equations.* We write

$$r = |r| e^{i\phi} \quad (73)$$

The variations of $|r_{\parallel}|$, $|r_{\perp}|$, ϕ_{\parallel} and ϕ_{\perp} are plotted in Figs 21.8 and 21.9 for $n_2/n_1 = 1.5$. The directions of the \mathbf{E} -vector in the reflected components are shown in Fig. 21.10.

Referring to Fig. 21.8 we note that when

$$\theta_1 = \theta_p \tan^{-1} \left(\frac{n_2}{n_1} \right) \approx 56^\circ, |r_{\parallel}| = 0$$

This is the Brewster's angle. At grazing incidence (i.e. as $\theta_1 \rightarrow 90^\circ$), both $|r_{\parallel}|$ and $|r_{\perp}|$ tend to 1 implying complete

*An alternative derivation of Fresnel equations has been given in Ref. 4, §33.6.

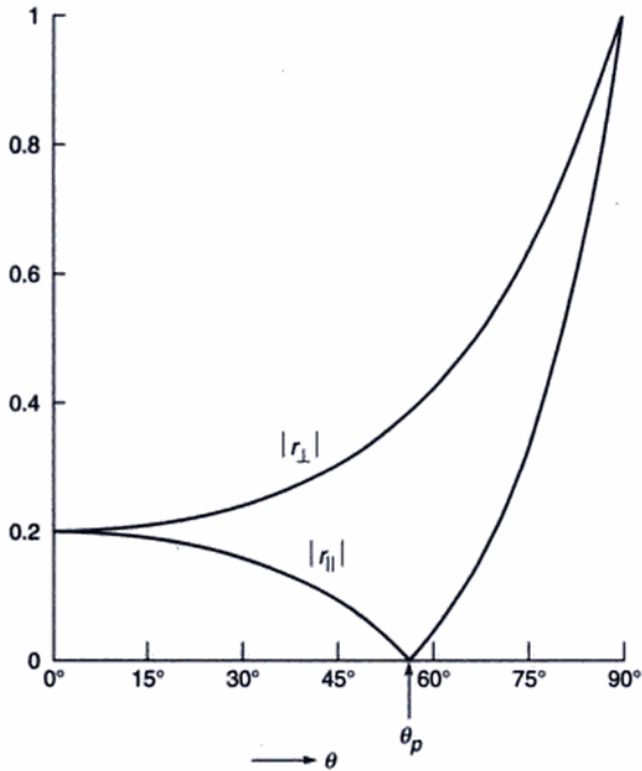


Fig. 21.8 Variation of $|r_{\parallel}|$ and $|r_{\perp}|$ with the angle of incidence when $n_2 = 1.5$ and $n_1 = 1.0$.

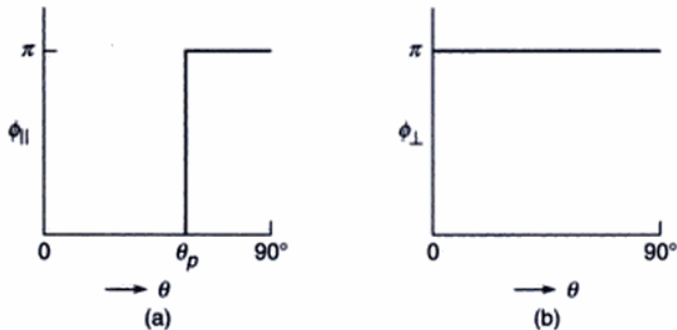


Fig. 21.9 The phase change on reflection (a) for the parallel component and (b) for the perpendicular component for $n_2 = 1.5$ and $n_1 = 1.0$; $\phi_{\perp} = \pi$ for all values of θ .

reflection. At normal incidence (i.e. $\theta_1 = 0$) any state of polarization can be thought of as parallel polarization or perpendicular polarization* and we should expect r_{\parallel} and r_{\perp} to give the same result. Figure 21.8 shows that both $|r_{\parallel}|$ and $|r_{\perp}|$ have the same value; however, Fig. 21.9 shows that whereas the perpendicular component predicts a phase change of π , there is no phase change associated with the parallel component. There is, however, no inconsistency, if

we study the direction of the electric vector associated with the reflected component [see (b) and (d) of Fig. 21.10].

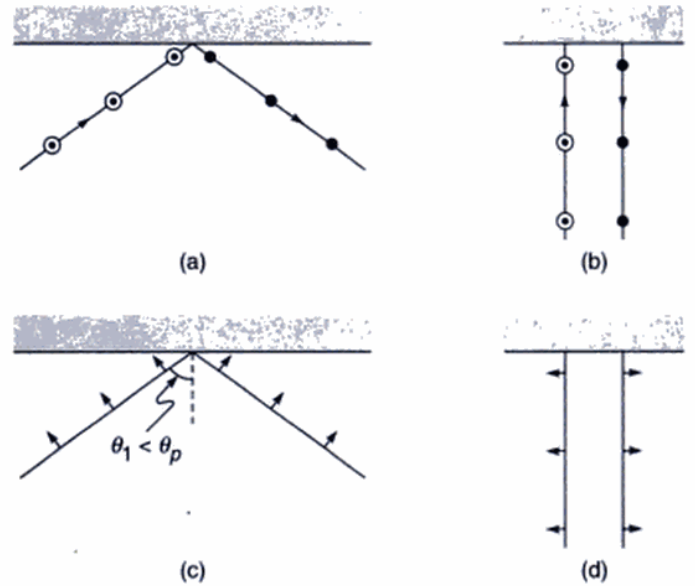


Fig. 21.10 For the perpendicular component, there is a phase change of π at all angles [(a) and (b)]. For the parallel component there is no phase change for $\theta_1 < \theta_p$ [see (c) and (d)]. Notice that at normal incidence, the electric field changes direction in both the cases.

We must now recapitulate. We first considered the case when the electric field associated with the incident wave was in the plane of incidence and assumed that the electric fields associated with the reflected and transmitted waves were also in the plane of incidence. Had we assumed that the reflection at the interface resulted in electric fields (E_{2y} and E_{3y}) along the y -direction associated with the transmitted and reflected waves, then the continuity of E_y and H_z at $x = 0$ would have given us

$$E_{3y} = E_{2y}$$

and

$$\frac{k_{3x} E_{3y}}{\omega \mu_0} = \frac{k_{2x} E_{2y}}{\omega \mu_0}$$

$$\Rightarrow -n_1 \cos \theta_1 E_{3y} = n_2 \cos \theta_2 E_{2y}$$

The above two equations would immediately result in the solution $E_{2y} = E_{3y} = 0$. Thus we may conclude that if the incident electric field lies in the plane of incidence then the electric fields associated with the reflected and transmitted waves must also lie in that plane. Similarly, if the incident

*This is due to the fact that at normal incidence the direction of propagation is coincident with the normal to the reflecting surface and any plane containing the normal could be thought of as the plane of incidence.

electric field is perpendicular to the plane of incidence then the electric fields associated with the reflected and transmitted waves will also lie perpendicular to the same plane. In general, for an arbitrary state of polarization of the incident wave, we must resolve the incident electric field in components which are parallel and perpendicular to the plane of incidence and consider the reflection (and transmission) of each of the components and then superpose to find the resultant state of polarization (see Example 21.9). Indeed by studying the polarization characteristics of the reflected wave, one can determine the (complex) refractive index of the material. This is known as the field of "ellipsometry"¹—a subject of profound importance. We will consider the reflection by a material of complex refractive index in Sec. 21.3.

Example 21.5 Let us consider the incidence of a plane electromagnetic wave on an air-glass interface (see Fig. 21.1). Thus $n_1 = 1.0$ and $n_2 = 1.5$ giving

$$\theta_p = \tan^{-1}(1.5) \approx 56.31^\circ$$

For $\theta_1 = 30^\circ$; $\theta_2 \approx 19.47^\circ$ we get

$$\begin{aligned} r_{\parallel} &\approx 0.1589; & t_{\parallel} &\approx 0.7725 \\ r_{\perp} &\approx -0.2404; & r_{\perp} &\approx 0.7596 \end{aligned}$$

On the other hand, for $\theta_1 = 89^\circ$ (grazing incidence), $\theta_2 = 41.80^\circ$ and

$$\begin{aligned} r_{\parallel} &\approx -0.9321 \quad (\sim 87\% \text{ reflection}); & t_{\parallel} &\approx 0.0452 \\ r_{\perp} &\approx -0.9693; & \text{and} & & t_{\perp} &\approx 0.0307 \end{aligned}$$

Example 21.6 We next consider the incidence of a plane electromagnetic wave on a rarer medium like a glass air interface. Thus $n_1 = 1.5$ and $n_2 = 1.0$ giving

$$\theta_p = \tan^{-1}\left(\frac{1}{1.5}\right) \approx 33.69^\circ \quad \text{and} \quad \theta_c = \sin^{-1}\left[\frac{1}{1.5}\right] \approx 41.81^\circ$$

(i) For $\theta_1 = 30^\circ$, $\theta_2 = 48.59^\circ$ and

$$\begin{aligned} r_{\parallel} &\approx -0.06788; & t_{\parallel} &\approx +1.3982 \\ r_{\perp} &\approx +0.3252; & t_{\perp} &\approx +1.3252 \end{aligned}$$

(ii) For $\theta_1 = 60^\circ$, $\cos \theta_2 = i\alpha$ with $\alpha = 0.82916$. Thus

$$\begin{aligned} r_{\parallel} &= \frac{n_2 \cos \theta_1 - n_1 \cos \theta_2}{n_2 \cos \theta_1 + n_1 \cos \theta_2} \\ &= \frac{0.5 - i1.5\alpha}{0.5 + i1.5\alpha} \\ &\approx -0.7217 - i0.6922 \\ &\approx e^{-0.7567\pi i} \end{aligned}$$

[Use of Eq. (53) would give the same result]

$$\begin{aligned} t_{\parallel} &= \frac{2n_1 \cos \theta_1}{n_2 \cos \theta_1 + n_1 \cos \theta_2} \\ &\approx \frac{1.5}{0.5 + i1.5\alpha} \approx 0.41739 - i1.0382 \\ &\approx 1.1190 e^{-0.3783\pi i} \\ r_{\perp} &= \frac{n_1 \cos \theta_1 - n_2 \cos \theta_2}{n_1 \cos \theta_1 + n_2 \cos \theta_2} \\ &= \frac{0.75 - i\alpha}{0.75 + i\alpha} \\ &\approx -0.1 - i0.9950 \approx e^{-0.532\pi i} \end{aligned}$$

(Notice that $|r_{\parallel}| = |r_{\perp}| = 1$)
and

$$\begin{aligned} t_{\perp} &= \frac{2n_1 \cos \theta_1}{n_1 \cos \theta_1 + n_2 \cos \theta_2} \\ &\approx 0.9 - i0.995 \\ &\approx 1.3416 e^{-0.266\pi i} \end{aligned}$$

Example 21.7 Consider a linearly polarized electromagnetic wave (with its electric vector along the y-direction of magnitude 5 V/m) propagating in vacuum. It is incident on a dielectric interface at $x = 0$ at an angle of incidence of 30° . The frequency associated with the wave is 6×10^{14} Hz. The refractive index of the dielectric is 1.5. Write the complete expressions for the electric and magnetic fields associated with the incident, reflected and transmitted waves.

Solution: The wave vector associated with the incident wave is given by

$$\begin{aligned} \mathbf{k}_1 &= (k_0 \cos 30^\circ) \hat{x} + (k_0 \sin 30^\circ) \hat{z} \\ &= \frac{\sqrt{3}}{2} k_0 \hat{x} + \frac{1}{2} k_0 \hat{z} \end{aligned}$$

Thus

$$\mathbf{E}_1 = \hat{y} 5 \exp \left[i \left(\frac{\sqrt{3}}{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \text{ V/m}$$

where

$$k_0 = \frac{2\pi}{\lambda_0} = 4\pi \times 10^6 \text{ m}^{-1}; \quad \omega = 12\pi \times 10^{14} \text{ Hz}$$

Now

$$\sin \theta_2 = \frac{n_1 \sin \theta_1}{n_2} = \frac{1}{3} \Rightarrow \cos \theta_2 = \frac{\sqrt{8}}{3}$$

Thus

$$\begin{aligned} r_{\perp} &= \frac{n_1 \cos \theta_1 - n_2 \cos \theta_2}{n_1 \cos \theta_1 + n_2 \cos \theta_2} = -0.2404 \\ \Rightarrow R_{\perp} &= 0.057796 \end{aligned}$$

and

$$t_{\perp} = \frac{2 \cos \theta_1 \sin \theta_2}{\sin (\theta_1 + \theta_2)} = 0.7596$$

implying

$$T_{\perp} = \frac{n_2 \cos \theta_2}{n_1 \cos \theta_1} |t_{\perp}|^2 = 0.942204$$

showing that $R_{\perp} + T_{\perp} = 1$. Now

$$\begin{aligned} \mathbf{k}_2 &= \hat{\mathbf{x}}(n_2 k_0 \cos \theta_2) + \hat{\mathbf{z}}(n_2 k_0 \sin \theta_2) \\ &= \hat{\mathbf{x}}(\sqrt{2} k_0) + \hat{\mathbf{z}}\left(\frac{1}{2} k_0\right) \end{aligned}$$

and

$$\begin{aligned} \mathbf{k}_3 &= -\hat{\mathbf{x}} k_0 \cos \theta_1 + \hat{\mathbf{z}}(k_0 \sin \theta_1) \\ &= -\hat{\mathbf{x}}\left(\frac{\sqrt{3}}{2} k_0\right) + \hat{\mathbf{z}}\left(\frac{1}{2} k_0\right) \end{aligned}$$

Thus the electric fields associated with the transmitted and reflected waves would be given by

$$\mathbf{E}_2 = 3.8 \hat{\mathbf{y}} \exp \left[i \left(\sqrt{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \text{ V/m}$$

and

$$\mathbf{E}_3 = -1.2 \hat{\mathbf{y}} \exp \left[i \left(-\frac{\sqrt{3}}{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \text{ V/m}$$

respectively. Notice that the values of k_z in \mathbf{E}_1 , \mathbf{E}_2 and \mathbf{E}_3 are the same [see Eq. (7)]. The corresponding magnetic fields can be calculated by using Eq. (59) to obtain

$$\begin{aligned} \mathbf{H}_1 &= 5 \frac{k_1}{\omega \mu_0} (-\hat{\mathbf{x}} \sin \theta_1 + \hat{\mathbf{z}} \cos \theta_1) \times \\ &\quad \exp \left[i \left(\frac{\sqrt{3}}{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \end{aligned}$$

$$\begin{aligned} \mathbf{H}_2 &= 3.8 \frac{k_2}{\omega \mu_0} (-\hat{\mathbf{x}} \sin \theta_2 + \hat{\mathbf{z}} \cos \theta_2) \times \\ &\quad \exp \left[i \left(\sqrt{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \end{aligned}$$

and

$$\begin{aligned} \mathbf{H}_3 &= -1.2 \frac{k_1}{\omega \mu_0} (-\hat{\mathbf{x}} \sin \theta_1 - \hat{\mathbf{z}} \cos \theta_1) \times \\ &\quad \exp \left[i \left(-\frac{\sqrt{3}}{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \end{aligned}$$

where

$$\frac{k_1}{\omega \mu_0} = \frac{k_0}{\omega \mu_0} = \frac{1}{c \mu_0} = \frac{1}{120 \pi} \text{ MKS units}$$

$$\text{and } \frac{k_2}{\omega \mu_0} = \frac{k_0 n_2}{\omega \mu_0} = \frac{n_2}{c \mu_0} = \frac{1}{80 \pi} \text{ MKS units}$$

Example 21.8 Consider once again the situation described in the above example except that the magnetic vector is now along the y-direction. Formulate the complete expressions for the electric fields associated with the incident, reflected and transmitted waves.

Solution: Referring to Fig. 21.1, we have

$$\begin{aligned} \mathbf{E}_1 &= 5 \left(\frac{1}{2} \hat{\mathbf{x}} - \frac{\sqrt{3}}{2} \hat{\mathbf{z}} \right) \times \\ &\quad \exp \left[i \left(\frac{\sqrt{3}}{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \text{ V/m} \end{aligned}$$

Now

$$r_{\parallel} = \frac{n_2 \cos \theta_1 - n_1 \cos \theta_2}{n_2 \cos \theta_1 + n_1 \cos \theta_2} = 0.1589$$

$$\Rightarrow R_{\parallel} = 0.02525$$

$$\text{and } t_{\parallel} = \frac{2 n_1 \cos \theta_1}{n_2 \cos \theta_1 + n_1 \cos \theta_2} = 0.7726$$

implying

$$T_{\parallel} = \frac{n_2 \cos \theta_2}{n_1 \cos \theta_1} |t_{\parallel}|^2 = 0.97475$$

showing that $R_{\parallel} + T_{\parallel} = 1$. Furthermore,

$$\begin{aligned} \mathbf{E}_2 &= 3.863 \left(\frac{1}{3} \hat{\mathbf{x}} - \frac{\sqrt{8}}{3} \hat{\mathbf{z}} \right) \times \\ &\quad \exp \left[i \left(\sqrt{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \text{ V/m} \end{aligned}$$

$$\begin{aligned} \mathbf{E}_3 &= 0.7945 \left(\frac{1}{2} \hat{\mathbf{x}} + \frac{\sqrt{3}}{2} \hat{\mathbf{z}} \right) \times \\ &\quad \exp \left[i \left(-\frac{\sqrt{3}}{2} k_0 x + \frac{1}{2} k_0 z - \omega t \right) \right] \text{ V/m} \end{aligned}$$

Example 21.9 For the situation described in Example 21.7, consider a right-circularly polarized wave incident at the air-glass interface at $\theta_1 = 30^\circ$. Determine the state of polarization of the reflected and transmitted fields.

Solution: We refer to Fig. 21.11. We must resolve the electric field in components parallel and perpendicular to the plane of incidence. Neglecting the space-dependent parts, if we write for the y-component of the incident field

$$E_{\perp} = E_y = E_0 \cos \omega t$$

then for the beam to be right-circularly polarized, the parallel component must be given by

$$E_{\parallel} = E_0 \cos \left(\omega t + \frac{\pi}{2} \right) = -E_0 \sin \omega t$$

and

$$t_{\perp} = \frac{E_{20}}{E_{10}} = 1 + r_{\perp} = \frac{2}{1 + \frac{\alpha + i\beta}{\omega\sqrt{\epsilon_1\mu_1}} \frac{\mu_1}{\mu_2}} \quad (79)$$

For a good conductor, $\frac{\sigma}{\epsilon\omega} \gg 1$ and

$$\alpha \approx \beta \approx \left(\frac{\omega\sigma\mu_2}{2} \right)^{1/2} \quad (80)$$

Thus

$$r_{\perp} = \frac{E_{30}}{E_{10}} \approx \frac{1 - (1+i)\Delta}{1 + (1+i)\Delta} \quad (81)$$

$$t_{\perp} = \frac{E_{20}}{E_{10}} \approx \frac{2}{1 + (1+i)\Delta} \quad (82)$$

where

$$\Delta = \left(\frac{\sigma\mu_1}{2\mu_2\epsilon_1\omega} \right)^{1/2} \quad (83)$$

For infinite conductivity, $\Delta \rightarrow \infty$ and

$$E_{30} = -E_{10} \quad E_{20} = 0 \quad (84)$$

showing that there is a phase change of π on reflection. Further, the energy is completely reflected and the field inside the conductor is identically zero.* For a finite (but large) value of σ , an approximate expression for the reflection coefficient can be obtained in the following manner:

$$\begin{aligned} R &= \left| \frac{E_{30}}{E_{10}} \right|^2 = \left| -\frac{1 - \frac{1}{(1+i)\Delta}}{1 + \frac{1}{(1+i)\Delta}} \right|^2 \\ &= \left| \left(1 - \frac{1}{(1+i)\Delta} \right) \left(1 - \frac{1}{(1+i)\Delta} \right) \right|^2 \\ &\approx \left| 1 - \frac{2}{(1+i)\Delta} \right|^2 \approx 1 - \frac{2}{\Delta} \\ &\approx 1 - 2 \left(\frac{2\mu_2\epsilon_1\omega}{\sigma\mu_1} \right)^{1/2} \end{aligned} \quad (85)$$

For non-magnetic media, with

$$\begin{aligned} \epsilon_1 &= \epsilon_0 = 8.854 \times 10^{-12} \text{ C}^2/\text{Nm}^2 \\ \omega &= 2\pi \times 10^{10} \text{ sec}^{-1}, \quad \sigma \approx 3 \times 10^7 \text{ mhos/m (silver)}, \\ R &\approx 0.9996. \end{aligned}$$

Thus about 99.96% of light is reflected. This is the reason why metals are such good reflectors. Notice that the reflection coefficient increases with decrease in frequency.

When the incidence is not normal one must substitute the following expression for $\cos \theta_2$.

$$\begin{aligned} \cos \theta_2 &= \sqrt{1 - \sin^2 \theta_2} = \sqrt{1 - (k_1/k_2)^2 \sin^2 \theta_1} \\ &= \left[1 - \frac{\omega^2 \epsilon_1 \mu_1}{(\alpha + i\beta)^2} \sin^2 \theta_1 \right]^{1/2} \\ &\approx 1 - \frac{1}{2} \frac{\omega^2 \epsilon_1 \mu_1 \sin^2 \theta_1}{\frac{\omega^2 \sigma^2 \mu_2^2}{4} (1+i)^2} \\ &\approx 1 \end{aligned}$$

The last expression being valid for good conductors. Thus for the transmitted wave we can write

$$\begin{aligned} \mathbf{E}_2 &= \mathbf{E}_{20} \exp [i(\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] \\ &= \mathbf{E}_{20} \exp [i\{k_2 \cos \theta_2 x + k_2 \sin \theta_2 z - \omega t\}] \\ &= \mathbf{E}_{20} \exp [i\{\alpha(1+i)x + k_1 \sin \theta_1 z - \omega t\}] \\ &\approx \mathbf{E}_{20} \exp(-\alpha x) \exp [i\{\alpha x + k_1 \sin \theta_1 z - \omega t\}] \end{aligned} \quad (86)$$

For a good conductor, $\alpha \gg k_1$ and the wave (having an amplitude exponentially decreasing in the x -direction) propagates along the x -axis.

21.4 REFLECTIVITY OF A DIELECTRIC FILM

In this section we will calculate the reflectivity of a dielectric film for a plane wave incident normally on it. We will determine the thickness of the film for which the film will become antireflecting and compare our results with those obtained in Sec. 13.4. In Problem 21.9, we will apply our results to a Fabry-Perot interferometer [cf. Sec. 14.2].

We consider a plane wave incident normally on a dielectric film of thickness d (see Fig. 21.12). Without any loss of generality, we assume the electric field to be along the y -axis. Thus the electric fields in media 1, 2 and 3 are given by

$$\left. \begin{aligned} \mathbf{E}_1 &= \hat{y} E_{10}^+ e^{i(k_1 x - \omega t)} + \hat{y} E_{10}^- e^{-i(k_1 x + \omega t)} \\ \mathbf{E}_2 &= \hat{y} E_{20}^+ e^{i(k_2 x - \omega t)} + \hat{y} E_{20}^- e^{-i(k_2 x + \omega t)} \\ \mathbf{E}_3 &= \hat{y} E_{30}^+ e^{i(k_3(x-d) - \omega t)} \end{aligned} \right\} \quad (87)$$

*It should be noted that if $\sigma \rightarrow \infty$ (i.e., for a perfect conductor) $r_{\parallel} \rightarrow +1$ and $r_{\perp} \rightarrow -1$ [see Eqs (46) and (75)] even for non-normal incidence. Thus, if a right-circularly polarized wave is incident on a perfect conductor then the reflected light will be left-circularly polarized.

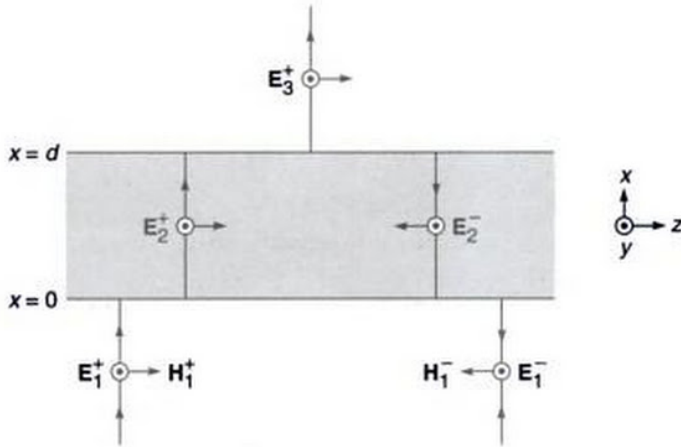


Fig. 21.12 Reflection of a plane wave incident normally on a dielectric slab of thickness d .

where E_{10}^+ and E_{10}^- represent the amplitudes of the forward and backward propagating waves in region 1; similarly for other fields. Since the 3rd medium extends to infinity, there is no backward propagating wave in region 3. For E_3 , we have for the sake of convenience, introduced a phase factor of $\exp[-ik_3d]$; this term makes the analysis more straightforward.

The corresponding magnetic field is given by [see Eq. (66) of the previous chapter]:

$$\mathbf{H} = \frac{\mathbf{k} \times \mathbf{E}}{\omega\mu} \quad (88)$$

where

$$\begin{aligned} \mathbf{k} &= k \hat{\mathbf{x}} \text{ for waves propagating in the } +x \text{ direction} \\ &= -k \hat{\mathbf{x}} \text{ for waves propagating in the } -x \text{ direction} \end{aligned}$$

Thus

$$\begin{aligned} \mathbf{H}_1 &= \hat{\mathbf{z}} \frac{k_1}{\omega\mu_0} [E_{10}^+ e^{i(k_1x - \omega t)} - E_{10}^- e^{-i(k_1x + \omega t)}] \\ \mathbf{H}_2 &= \hat{\mathbf{z}} \frac{k_2}{\omega\mu_0} [E_{20}^+ e^{i(k_2x - \omega t)} - E_{20}^- e^{-i(k_2x + \omega t)}] \\ \mathbf{H}_3 &= \hat{\mathbf{z}} \frac{k_3}{\omega\mu_0} E_{30}^+ e^{i[k_3(x-d) - \omega t]} \end{aligned} \quad (89)$$

Both E_y and H_z represent tangential components, and should therefore be continuous at interfaces $x = 0$ and $x = d$. The continuity conditions at $x = 0$ give us

$$E_{10}^+ + E_{10}^- = E_{20}^+ + E_{20}^-$$

and

$$\frac{k_1}{\omega\mu_0} (E_{10}^+ - E_{10}^-) = \frac{k_2}{\omega\mu_0} (E_{20}^+ - E_{20}^-)$$

or

$$E_{10}^+ - E_{10}^- = \frac{n_2}{n_1} (E_{20}^+ - E_{20}^-)$$

where we have used the relations

$$k_2 = \frac{\omega}{c} n_2 \text{ and } k_1 = \frac{\omega}{c} n_1$$

Simple manipulations give us

$$\begin{pmatrix} E_{10}^+ \\ E_{10}^- \end{pmatrix} = \begin{pmatrix} \frac{n_1 + n_2}{2n_1} & \frac{n_1 - n_2}{2n_1} \\ \frac{n_1 - n_2}{2n_1} & \frac{n_1 + n_2}{2n_1} \end{pmatrix} \begin{pmatrix} E_{20}^+ \\ E_{20}^- \end{pmatrix} \quad (90)$$

Similarly, the continuity of E_y and H_z at $x = d$ give us

$$E_{20}^+ e^{i\delta} + E_{20}^- e^{-i\delta} = E_{30}^+$$

$$E_{20}^+ e^{i\delta} - E_{20}^- e^{-i\delta} = \frac{n_3}{n_2} E_{30}^+$$

where $\delta = k_2 d$. Elementary manipulations give us

$$\begin{pmatrix} E_{20}^+ \\ E_{20}^- \end{pmatrix} = \begin{pmatrix} \frac{n_2 + n_3}{2n_2} e^{-i\delta} \\ \frac{n_2 - n_3}{2n_2} e^{i\delta} \end{pmatrix} E_{30}^+ \quad (91)$$

Combining Eqs (90) and (91), we get

$$\begin{aligned} E_{10}^+ &= \left[\left(\frac{n_1 + n_2}{2n_1} \right) \left(\frac{n_2 + n_3}{2n_2} e^{-i\delta} \right) \right. \\ &\quad \left. + \left(\frac{n_1 - n_2}{2n_1} \right) \left(\frac{n_2 - n_3}{2n_2} e^{i\delta} \right) \right] E_{30}^+ \end{aligned} \quad (92)$$

and

$$\begin{aligned} E_{10}^- &= \left[\left(\frac{n_1 - n_2}{2n_1} \right) \left(\frac{n_2 + n_3}{2n_2} e^{-i\delta} \right) \right. \\ &\quad \left. + \left(\frac{n_1 + n_2}{2n_1} \right) \left(\frac{n_2 - n_3}{2n_2} e^{i\delta} \right) \right] E_{30}^+ \end{aligned} \quad (93)$$

Dividing Eq. (92) by Eq. (93), we get the amplitude reflection coefficient

$$r = \frac{E_{10}^-}{E_{10}^+} = \frac{r_1 e^{-i\delta} + r_2 e^{i\delta}}{e^{-i\delta} + r_1 r_2 e^{i\delta}} \quad (94)$$

where

$$r_1 = \frac{n_1 - n_2}{n_1 + n_2} \quad (95)$$

and

$$r_2 = \frac{n_2 - n_3}{n_2 + n_3} \quad (96)$$

represent the Fresnel reflection coefficients at the first and second interface respectively. The reflectivity would therefore be given by

$$R = |r|^2 = \frac{r_1^2 + r_2^2 + 2r_1 r_2 \cos 2\delta}{1 + r_1^2 r_2^2 + 2r_1 r_2 \cos 2\delta} \quad (97)$$

In Sec. 13.4, we had discussed the above equation in detail. We should mention here that a more general analysis shows that the above equation remains valid even for oblique incidence with δ now equal to $k_2 d \cos \theta_2$, θ_2 being the angle of refraction in the second medium and r_1 and r_2 representing the appropriate Fresnel reflection coefficients corresponding to the particular angle of incidence and state of polarization. (The definition of δ here differs by a factor of 2 from the definition of δ in Chapter 13).

SUMMARY

- Consider the incidence of a linearly polarized electromagnetic wave on an interface of two dielectrics (which we assume to be $x = 0$); the $x - z$ plane is assumed to be the plane of incidence. Let $n_1 = \sqrt{\frac{\epsilon_1}{\epsilon_0}}$

and $n_2 = \sqrt{\frac{\epsilon_2}{\epsilon_0}}$ be the refractive indices of the two media. The incident wave, refracted wave and reflected waves can be written as

$$\begin{aligned} \mathbf{E}_1 &= \mathbf{E}_{10} \exp [i (\mathbf{k}_1 \cdot \mathbf{r} - \omega t)] && \text{incident wave} \\ \mathbf{E}_2 &= \mathbf{E}_{20} \exp [i (\mathbf{k}_2 \cdot \mathbf{r} - \omega t)] && \text{refracted wave} \\ \mathbf{E}_3 &= \mathbf{E}_{30} \exp [i (\mathbf{k}_3 \cdot \mathbf{r} - \omega t)] && \text{reflected wave} \end{aligned}$$

where \mathbf{E}_{10} , \mathbf{E}_{20} and \mathbf{E}_{30} are independent of space and time and

$$k_1 = \frac{\omega}{c} n_1 = k_3; \quad k_2 = \frac{\omega}{c} n_2$$

$$k_1 \sin \theta_1 = k_2 \sin \theta_2 = k_3 \sin \theta_3$$

where θ_1 , θ_2 and θ_3 are the angle of incidence, angle of refraction and angle of reflection respectively. The above equations readily give

$$n_1 \sin \theta_1 = n_2 \sin \theta_2 \text{ (Snell's law)}$$

and $\theta_1 = \theta_3$.

- For \mathbf{E}_1 lying in the $x - z$ plane (which is the plane of incidence)

$$\begin{aligned} \mathbf{E}_{10} &= E_{10} (\hat{x} \sin \theta_1 - \hat{z} \cos \theta_1) \\ \mathbf{E}_{20} &= t_{\parallel} E_{10} (\hat{x} \sin \theta_2 - \hat{z} \cos \theta_2) \\ \mathbf{E}_{30} &= r_{\parallel} E_{10} (\hat{x} \sin \theta_1 + \hat{z} \cos \theta_1) \end{aligned}$$

$$r_{\parallel} = \frac{n_2 \cos \theta_1 - n_1 \cos \theta_2}{n_2 \cos \theta_1 + n_1 \cos \theta_2} = \frac{\tan(\theta_1 - \theta_2)}{\tan(\theta_1 + \theta_2)}$$

$$\begin{aligned} t_{\parallel} &= \frac{2n_1 \cos \theta_1}{n_2 \cos \theta_1 + n_1 \cos \theta_2} \\ &= \frac{2 \cos \theta_1 \sin \theta_2}{\sin(\theta_1 + \theta_2) \cos(\theta_1 - \theta_2)} \end{aligned}$$

Notice that $r_{\parallel} = 0$ when $\theta_1 + \theta_2 = \frac{\pi}{2}$ implying

$$\theta_1 = \theta_p = \tan^{-1}(n_2/n_1).$$

This is the Brewster's angle.

- For \mathbf{E}_1 perpendicular to the plane of incidence (i.e., along \hat{y}),

$$\mathbf{E}_{10} = E_{10} \hat{y}; \quad \mathbf{E}_{20} = t_{\perp} E_{10} \hat{y} \quad \text{and} \quad \mathbf{E}_{30} = r_{\perp} E_{10} \hat{y}$$

with

$$r_{\perp} = \frac{n_1 \cos \theta_1 - n_2 \cos \theta_2}{n_1 \cos \theta_1 + n_2 \cos \theta_2} = -\frac{\sin(\theta_1 - \theta_2)}{\sin(\theta_1 + \theta_2)}$$

and

$$t_{\perp} = \frac{2n_1 \cos \theta_1}{n_1 \cos \theta_1 + n_2 \cos \theta_2} = \frac{2 \sin \theta_2 \cos \theta_1}{\sin(\theta_1 + \theta_2)}$$

- In both cases, if $n_2 < n_1$ and $\theta_1 > \theta_c = \sin^{-1}\left(\frac{n_2}{n_1}\right)$ we

have total internal reflection. We can still use the above expressions for r_{\parallel} , t_{\parallel} , r_{\perp} and t_{\perp} but we must remember that

$$\sin \theta_2 = \frac{n_1}{n_2} \sin \theta_1 \text{ will be greater than 1}$$

$$\text{and} \quad \cos \theta_2 = \sqrt{1 - \sin^2 \theta_2} = i\alpha$$

will be pure imaginary. Thus r_{\parallel} , t_{\parallel} , r_{\perp} and t_{\perp} will be complex quantities with $|r_{\parallel}| = 1 = |r_{\perp}|$ showing that the entire energy is reflected; however, there will be an evanescent wave in the second medium whose field will decay along the x -axis and propagate along the z -axis.

PROBLEMS

- 21.1 Show that in the limit of $\theta_1 \rightarrow 0$ (i.e., at normal incidence) the reflection coefficient is the same for parallel and perpendicular polarizations.
- 21.2 Consider a magnetic dielectric with a permeability such that $\mu/\mu_0 = \epsilon/\epsilon_0$. Show that for such a material the reflection coefficient for normal incidence is identically equal to zero. This realization is equivalent to the situation where the impedance is matched at the junction of two transmission lines.

(The quantity $\sqrt{\mu/\epsilon}$ can be considered as the intrinsic impedance of the medium.)

- 21.3 A right-circularly polarized beam is incident on a perfect conductor at 45° . Show that the reflected beam is left-circularly polarized.

- 21.4 Assume $n_1 = 1.5$ and $n_2 = 1.0$ (see Example 21.6)

(a) for $\theta_1 = 45^\circ$ show that

$$r_{\parallel} = +0.28 - i0.96; t_{\parallel} = 1.92 - i1.44$$

Similarly calculate r_{\perp} and t_{\perp} .

(b) On the other hand, for $\theta_1 = 33.69^\circ$ show that

$$r_{\parallel} = 0, t_{\parallel} = 1.5$$

$$r_{\perp} = +0.3846, t_{\perp} = 1.3846$$

- 21.5 Consider a right-circularly polarized beam incident on a medium of refractive index 1.6 at an angle of 60° . Calculate r_{\parallel} and r_{\perp} and show that the reflected beam is right elliptically polarized with its major axis much longer than its minor axis. What will happen at 58° ?

[Ans. $r_{\parallel} = -0.0249$, $r_{\perp} = -0.4581$]

- 21.6 Consider a y-polarized wave incident on a glass-air interface ($n_1 = 1.5$, $n_2 = 1.0$) at $\theta_1 = 45^\circ$ and at $\theta_1 = 80^\circ$. Write the complete expressions for the transmitted field and show that in the latter case it is an evanescent wave with depth of penetration ($= 1/\beta$) equal to about 8.8×10^{-8} m; assume $\lambda = 6000 \text{ \AA}$.

- 21.7 For gold, at $\lambda_0 = 6530 \text{ \AA}$ the complex refractive index is given by $n_2 = 0.166 + 3.15i$. Calculate k_2 and show that the reflectivity at normal incidence is

approximately 94%. [Hint: Use Eq. (75) directly]. On the other hand, at $\lambda_0 = 4000 \text{ \AA}$, $n_2 = 1.658 + 1.956i$; show that the reflectivity is only 39%].

- 21.8 Show that for $\delta = 0$, Eq. (97) takes the form

$$R = \left(\frac{n_1 - n_3}{n_1 + n_3} \right)^2 \quad (98)$$

as it indeed should be.

- 21.9 Using the various equations in Sec. 21.4 calculate the transmittivity and show that

$$T = \frac{\frac{1}{2} n_3 |E_3^+|^2}{\frac{1}{2} n_1 |E_1^+|^2} = 1 - R$$

- 21.10 Assume the third medium in Fig. 21.12 to be identical to the first medium, i.e., $n_3 = n_1$. Thus

$$r_2 = -r_1 = -\frac{n_1 - n_2}{n_1 + n_2}$$

Using Eq. (97), show that

$$R = \frac{F \sin^2 \delta}{1 + F \sin^2 \delta} \quad (99)$$

$$\text{where } F = \frac{4r_1^2}{(1 - r_1^2)^2} \quad (100)$$

is called the coefficient of finesse. Equation (99) is identical to the result derived in Sec. 14.2 while discussing the theory of Fabry-Perot interferometer. The definition of δ here differs by a factor of 2 from the definition of δ in Chapter 13.

REFERENCES AND SUGGESTED READINGS

1. J.M. Bennett and H.E. Bennett, 'Polarization' in *Handbook of Optics* (Ed. W.J. Driscoll), McGraw-Hill, New York, 1978.
2. O. Bryngdahl, 'Evanescent Waves in Optical Imaging', *Progress in Optics* (Ed. E. Wolf), Vol. XI, North-Holland, Amsterdam, 1973.
3. D.R. Corson and P. Lorrain, *Introduction to Electromagnetic Fields and Waves*, W.H. Freeman and Co., San Francisco, 1962.
4. R.P. Feynman, R.B. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Vol. I, Addison-Wesley, Reading, Mass., 1964.
5. A. Ghatak, and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, 1989. [Reprinted by Foundation Books, New Delhi].
6. J.R. Heitzler, 'The Largest Electromagnetic Waves', *Scientific American*, Vol. 206, 128, September 1962.
7. E.C. Jordan and K.G. Balmain, *Electromagnetic Waves and Radiating Systems*, Prentice-Hall, N.J., USA 1970.
8. W.K.H. Panofsky and M. Phillips, *Classical Electricity and Magnetism*, Addison-Wesley, Reading, Mass., 1962.
9. J.R. Reitz and F.J. Milford, *Foundations of Electromagnetic Theory*, Addison-Wesley, Reading, Mass., 1962.
10. H.S. Sandhu and G.B. Friemann, 'Change of Phase on Reflection', *American Journal of Physics*, Vol. 39, 388, 1971.

PART 6

Photons

This part consists of only one chapter namely on the particle model of radiation. The photoelectric effect (discovered by Hertz in 1888) had certain peculiarities which cannot be explained on the basis of wave theory. In 1905, Einstein provided a simple explanation of the peculiarities by assuming that light consisted of quanta of energy $h\nu$ (where ν is the frequency) and that the emission of a photoelectron was the result of the interaction of a single quantum (i.e. of the photon) with an electron. – for which Einstein received the 1921 Nobel Prize in Physics. The chapter also discusses the Compton effect (for which Professor Compton received the 1927 Nobel Prize in Physics) which established that the photon has a momentum equal to h/λ .

Chapter 22

The Particle Nature of Radiation

Are not the rays of light very small bodies emitted from shining substance?

Isaac Newton in OPTICKS

It is undeniable that there is an extensive group of data concerning radiation which shows that light has certain fundamental properties that can be understood much more readily from the standpoint of the Newton emission (particle) theory than from the standpoint of the wave theory. It is my opinion, therefore, that the next phase of the development of theoretical physics will bring us a theory of light that can be interpreted as a kind of fusion of the wave and emission theories.

—Albert Einstein (1909)

Important Milestones

- | | |
|------|---|
| 1637 | Using corpuscular model of light, Descartes derived Snell's law. |
| 1887 | Heinrich Hertz discovered the photoelectric effect in which a metal surface irradiated by light beam would emit electrons if the frequency of the radiation was above a certain critical value. |
| 1900 | In order to derive the blackbody radiation formula, Planck made a drastic assumption that the oscillators can only assume discrete energies. |
| 1905 | In a paper entitled <i>On a heuristic point of view about the creation and conversion of light</i> , Einstein introduced the light quanta. In this paper he wrote that for explanation of phenomena like blackbody radiation, production of electrons by ultraviolet light (which is the photoelectric effect) it is necessary to assume that 'when a light ray starting from a point is propagated, the energy is not continuously distributed over an ever increasing volume, but it consists of a finite number of energy quanta, localized in space, which move without being divided and which can be absorbed or emitted only as a whole'. Einstein received the 1921 Nobel prize in Physics for his <i>discovery of the law of photoelectric effect</i> and <i>not</i> for his theory of relativity. |
| 1923 | Compton reported his studies on the scattering of X-rays by solid materials (mainly graphite) and showed that the shift of the wavelength of the scattered photon could be explained by assuming the photon having momentum equal to h/λ . Compton received the 1927 Nobel prize in Physics for his <i>discovery of the effect named after him</i> . |
| 1926 | Gilbert Lewis, an American chemist, coined the word 'photon' to describe Einstein's 'localized energy quanta'. |

22.1 INTRODUCTION

In the earlier chapters, we have discussed the interference, diffraction and polarization of light. All these phenomena can be explained satisfactorily on the basis of the wave theory of light. We have also discussed the electromagnetic character of light waves (see Chapters 19 and 20) and have shown that the electromagnetic theory can be successfully used to explain the origin of refractive index (see Chapter 6), the phenomenon of double refraction (see Chapter 19)

and many other experimental results. However, there exist a large number of experimental phenomena that can only be explained on the basis of the corpuscular nature of radiation. In this chapter we will discuss the famous experiments on the photoelectric effect and the Compton effect which establish the particle nature of light—a wave model is totally inadequate to explain these effects; in Chapter 1, we had briefly discussed how one can reconcile to the dual nature of radiation (i.e., the wave and the particle aspects) on the basis of the quantum theory.

22.2 THE PHOTOELECTRIC EFFECT

In 1887, Hertz discovered that a metal irradiated by a light beam would emit electrons. These electrons are known as photoelectrons and can be collected by a metal plate P_2 as shown in Fig. 22.1. The photoelectrons constitute a current between the plates P_1 and P_2 which can be detected by means of an ammeter A . When the voltage across the plates is varied, the current also varies; typical variations of the current with voltage are shown in Fig. 22.2. The figure corresponds to monochromatic light of a particular wavelength and different curves correspond to different intensities of the beam. From the figure one can draw the following conclusions:

- (i) At zero voltage there is a finite value of the current implying that some of the emitted photoelectrons reach the metal surface P_2 .
- (ii) As the voltage is increased, the current increases till it reaches a saturation value; this will happen when the plate P_2 collects all the emitted photoelectrons.
- (iii) If the plate P_2 is kept at a slightly negative potential, there is a weak current implying that some of the

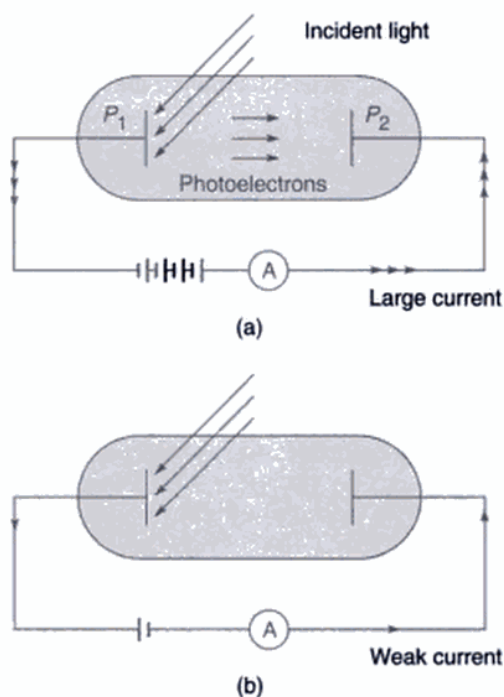


Fig. 22.1 If light (of certain frequency) is allowed to fall on a metal like sodium, electrons are emitted which can be collected by the plate P_2 . (a) and (b) correspond to positive and negative voltage applied to the plate P_2 . Even when the plate is kept at a low negative voltage, one can detect a small current.

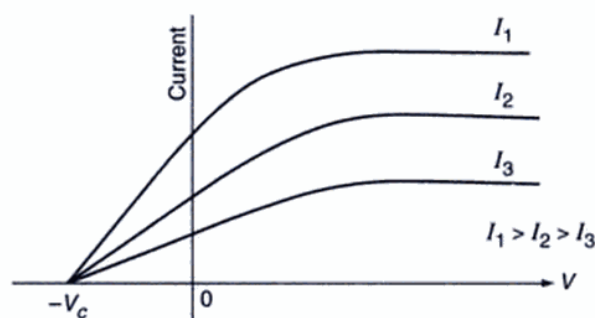


Fig. 22.2 Typical variation of the photocurrent with voltage. The curves correspond to light (of the same frequency) having different intensities.

photoelectrons do manage to reach the plate P_2 . However, beyond a certain voltage (which is shown as $-V_c$ in the figure) the current is zero; V_c is known as the cutoff voltage and the quantity $|q|V_c$ will represent the maximum kinetic energy of the photoelectrons (q represents the charge of the electron). For example, for sodium $V_c \approx 2.3$ Volts and for copper $V_c \approx 4.7$ Volts.

- (iv) If we do not change the wavelength of the incident radiation but make it more intense, the magnitude of the current will become larger as shown in Fig. 22.2 implying a greater emission of photoelectrons. Notice that the value of the cutoff potential remains the same; this important result implies that the maximum kinetic energy of the emitted photoelectrons does not depend on the intensity of the incident radiation.
- (v) If the frequency of the incident radiation is increased then the cutoff potential and hence the maximum kinetic energy of the electron ($= |q| V_c$) varies linearly with the frequency as shown in Fig. 22.3. Further, for frequencies less than a critical value (shown as ν_c in Fig. 22.3), there is *no* emission of photoelectrons no matter what the intensity of the incident radiation may be.

At first sight it appears that since electromagnetic waves carry energy, the wave model for light should be able to explain the emission of photoelectrons from a metal surface. However, there are certain peculiarities associated with photoelectric effect, which cannot be satisfactorily explained by means of a wave model:

1. The first peculiarity is the fact that the maximum kinetic energy of the electrons does not depend on the intensity of the incident radiation, it only depends on its frequency; further, a greater intensity leads to a larger number of electrons constituting a larger

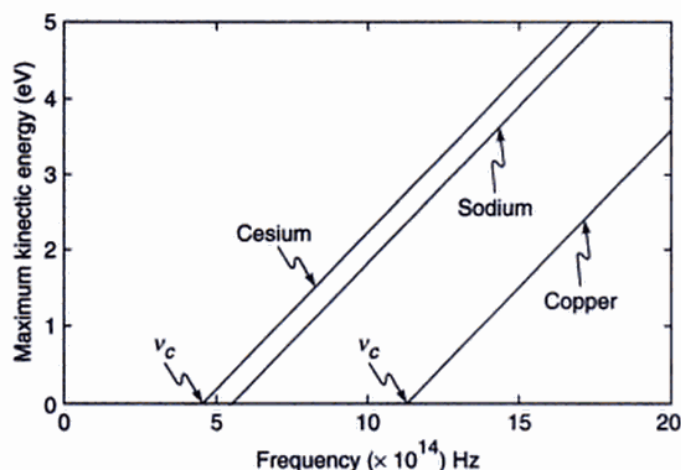


Fig. 22.3 The variation of the maximum kinetic energy of the electrons as a function of frequency of the incident light.

current. Thus, a faint violet light would eject electrons of greater kinetic energy than an intense yellow light although the latter would produce a large number of electrons. A wave model would, however, predict that a large intensity of the incident radiation would result in a greater kinetic energy of the emitted electrons.

- The second peculiarity is the fact that there is almost no time lag between the times of incidence of the radiation and the ejection of the photoelectron. For weak intensities of the incident beam the wave theory predicts considerable time lag for the electrons to absorb enough energy to leave the metal surface. This can be illustrated by considering a specific example. One can observe a detectable photocurrent if the surface of sodium metal is illuminated by violet light of intensity as low as 10^{-10} W/cm². Now, ten layers of sodium will contain about

$$\frac{6 \times 10^{23} \times 10 \times 10^{-8}}{23} \approx 2 \times 10^{15} \text{ atoms/cm}^2$$

where we have assumed the density of sodium to be $\approx 1\text{g/cm}^3$. Assuming that the energy is uniformly absorbed by the upper ten layers of sodium, each atom would receive energy at the rate of

$$\frac{10^{-10}}{2 \times 10^{15}} \approx 5 \times 10^{-26} \text{ J/s} \approx 3 \times 10^{-7} \text{ eV/s}$$

Assuming that an electron should acquire an energy ~ 1 eV to escape from the metal, we should expect a time lag of order 10^7 sec (\sim few months). However, the experiments show that there is no detectable time lag between the incidence of the radiation and the emission of the photoelectrons. Indeed, in 1928, Lawrence and Beams had devised an experiment to

find out whether the time lag was $\leq 3 \times 10^{-9}$ sec; the experiment gave a negative result.

In 1905, Einstein provided a simple explanation of the above mentioned peculiarities. He argued that light consisted of quanta of energy $h\nu$ (where ν is the frequency) and that the emission of a photoelectron was the result of the interaction of a single quantum (i.e., of the photon) with an electron. In his 1905 paper³, Einstein wrote:

Monochromatic radiation behaves as if it consists of mutually independent energy quanta of magnitude $[h\nu]$.

Einstein's theory gives a very satisfactory explanation of the photoelectric effect. According to this theory, a light beam (of frequency ν) essentially consists of individual corpuscles called photons; we may mention here that it was in 1926 that Gilbert Lewis, an American chemist, coined the word 'photon' to describe Einstein's 'localized energy quanta'. Each photon carries an energy equal to $h\nu$. This corpuscular model can explain all the observations discussed above. Thus, for all frequencies below the cutoff ν_c , each photon will carry energy less than $h\nu_c$ which will not be sufficient to eject the electron from the metal. For $\nu > \nu_c$, a major fraction of the excess energy $[= h(\nu - \nu_c)]$ appears as kinetic energy of the emitted electron. Further, the non-measurable time lag between the incidence of the radiation and the ejection of the electron follows immediately from the corpuscular nature of the radiation. Indeed, the observed maximum kinetic energy of the photoelectrons is linearly related to the frequency of the incident radiation and one may write (see Fig. 22.3)

$$T_{\max} = -B + h\nu = h(\nu - \nu_c) \quad (1)$$

where $B(= h\nu_c)$ is a constant and h is the Planck's constant ($= 6.627 \times 10^{-27}$ erg-sec). The frequency ν_c represents the cutoff frequency and is a characteristic of the metal. For example,

$$\text{for cesium } B \approx 1.9 \text{ eV} \Rightarrow \nu_c \approx 4.6 \times 10^{14} \text{ Hz}$$

$$\text{for sodium } B \approx 2.3 \text{ eV} \Rightarrow \nu_c \approx 5.6 \times 10^{14} \text{ Hz}$$

$$\text{for copper } B \approx 4.7 \text{ eV} \Rightarrow \nu_c \approx 11.4 \times 10^{14} \text{ Hz}$$

In Fig. 22.3, ν_c is the intercept on the horizontal axis. In 1909, Einstein wrote¹

It is undeniable that there is an extensive group of data concerning radiation which shows that light has certain fundamental properties that can be understood much more readily from the standpoint of the Newton emission (particle) theory than from the standpoint of the wave theory. It is my opinion, therefore, that the next phase of the development of theoretical physics will bring us a theory of light that can be interpreted as a kind of fusion of the wave and emission theories.

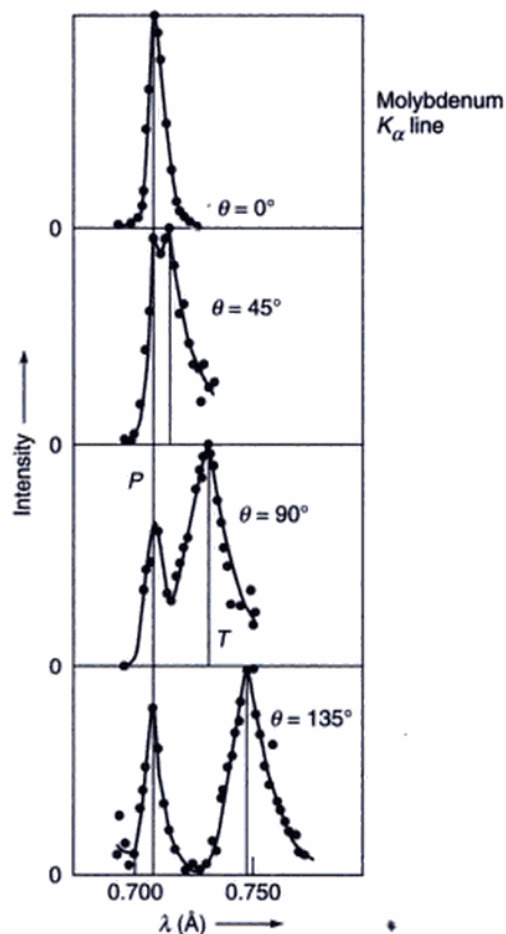


Fig. 22.7 The intensity variation as a function of the wavelength of the scattered photon. The vertical line (marked *P*) corresponds to the unmodified wavelength $\lambda = 0.711 \text{ \AA}$. The second vertical line (marked *T*) corresponds to the wavelength as predicted by Eq. (5). The figure has been adapted from the original paper of Compton (Ref. 10).

K_{α} line ($\lambda = 0.711 \text{ \AA}$). The sample used was graphite. Notice that at each value of θ , there are two peaks; the first peak appears at almost the same wavelength as the primary beam. This peak is because of the fact that the photon may be scattered by the whole atom; consequently, the quantity m_0 appearing in Eq. (4) is not the electron mass but the mass of the carbon atom (which is about 22,000 times that of the mass of the electron). Thus the wavelength shift is negligible. The second peak corresponds to the Compton shift. In each figure, the two vertical lines correspond to the unmodified wavelength and the modified wavelength as given by Eq. (5) and one can see a good agreement between the predicted and observed values.

Further evidence of the validity of the above theory was provided by the experiments carried out by Compton and

Simon who studied the scattering of X-rays through super-saturated water vapour. In the scattering process, the recoil electrons formed tracks of condensed droplets; however, the light quantum did not leave any track. Now, if the light quantum undergoes another Compton scattering then from the track of the second recoil electron one can determine the path of the light quantum by simply joining the line of the starting points of the two recoil electrons. Although there was considerable uncertainty in the analysis of the experimental data (because of the presence of many tracks) Compton and Simon could establish agreement between theoretical results and experimental data.

22.3.1 Kinematics of Compton Scattering

We next consider the scattering of a photon by an electron as shown in Fig. 22.4. The scattered photon is assumed to have a frequency ν' . Conservation of energy leads to

$$h\nu = h\nu' + E_k \quad (6)$$

where E_k represents the kinetic energy imparted to the electron. Conserving the x and y components of the momentum, we have

$$\frac{h\nu}{c} = \frac{h\nu'}{c} \cos\theta + p \cos\phi \quad (7)$$

and

$$0 = \frac{h\nu'}{c} \sin\theta - p \sin\phi \quad (8)$$

where p represents the momentum of the electron after collision, θ and ϕ represent the angles made by the scattered photon and the electron with the original direction of the photon (see Fig. 22.4). It will be shown that for a measurable Compton's effect, the frequency ν should be in the X-ray or in the γ -ray region (for X-rays $\lambda \leq 1 \text{ \AA}$ and $h\nu \geq 10^4 \text{ eV}$). For such high energy photons, the velocity imparted to the electron is comparable to the speed of light and one must use proper relativistic expressions for E_k and p . Now, according to the theory of relativity, the kinetic energy E_k of the scattered electron would be given by

$$E_k = E - m_0c^2 = mc^2 - m_0c^2 = \frac{m_0c^2}{\sqrt{1-\beta^2}} - m_0c^2 \quad (9)$$

where $\beta = v/c$, m_0 represents the rest mass of the electron, v the speed of the electron and c the speed of light in free-space; the quantities E and m_0c^2 are known as the total energy and the rest mass energy of the electron. Further, the relativistic momentum of the electron is given by

$$p = mv = \frac{m_0v}{\sqrt{1-\beta^2}} \quad (10)$$

Now,

$$\begin{aligned} p^2 c^2 + m_0^2 c^4 &= \frac{m_0^2 v^2 c^2}{1 - v^2/c^2} + m_0^2 c^4 \\ &= \frac{m_0^2 c^4}{1 - v^2/c^2} = m^2 c^4 \end{aligned}$$

or

$$\begin{aligned} p^2 c^2 + m_0^2 c^4 &= E^2 = (E_k + m_0 c^2)^2 \\ &= E_k^2 + m_0^2 c^4 + 2E_k m_0 c^2 \end{aligned}$$

Thus,

$$E_k^2 + 2E_k m_0 c^2 = p^2 c^2$$

Substituting for E_k from Eq. (6), we get

$$h^2 (v - v')^2 + 2h(v - v') m_0 c^2 = p^2 c^2 \quad (11)$$

Further, Eqs (7) and (8) can be rewritten in the form

$$p \cos \phi = \frac{hv}{c} - \frac{hv'}{c} \cos \theta \quad (12)$$

and

$$p \sin \phi = \frac{hv'}{c} \sin \theta \quad (13)$$

In order to eliminate ϕ , we square and add to obtain

$$p^2 = \left(\frac{hv}{c}\right)^2 + \left(\frac{hv'}{c}\right)^2 - \frac{2h^2 vv'}{c^2} \cos \theta \quad (14)$$

Substituting in Eq. (11), we obtain

$$\begin{aligned} h^2(v^2 - 2vv' + v'^2) + 2h(v - v') m_0 c^2 \\ = h^2 v^2 + h^2 v'^2 - 2h^2 vv' \cos \theta \end{aligned}$$

or

$$\frac{2h(v - v') m_0 c^2}{2vv'} = h^2 (1 - \cos \theta)$$

or

$$\Delta \lambda = \lambda' - \lambda = \frac{h}{m_0 c} (1 - \cos \theta)$$

or

$$\Delta \lambda = \frac{2h}{m_0 c} \sin^2 \frac{\theta}{2} \quad (15)$$

which gives us the Compton shift*.

22.4 THE PHOTON MASS

Because of the fact that the photon has energy ($= hv/c$) we may assume it to have an inertial mass given by

$$m = \frac{hv}{c^2} \quad (16)$$

Thus when a light beam passes near a heavy star, its trajectory ought to get deflected. Indeed, the light coming from a distant star does get slightly deflected when passing near the sun, which has been experimentally observed.

Also, we may expect that when a photon leaves a star, its energy should decrease because of the gravitation field. This indeed happens and manifests itself in decrease in frequency which is usually referred as the *gravitational red shift*. One can approximately calculate the red shift by noting that the potential energy on the surface of the star would be

$$V \approx -\frac{GMm}{R} = -\frac{GM}{R} \cdot \frac{hv}{c^2} \quad (17)$$

where M is the mass of the star, R its radius and G the Gravitational constant. Thus when the light beam reaches earth its frequency would become

$$hv' = hv - \frac{GM}{R} \frac{hv}{c^2}$$

or

$$\frac{\Delta v}{v} = \frac{v - v'}{v} = \frac{GM}{Rc^2} \quad (18)$$

(we have neglected the effect of the earth's gravitational field). From the above equation, we see that if the mass of the star is so large so that the RHS exceeds unity then the light beam will not be able to escape from the star—this is a black hole. We should mention that in discussing black holes, we must use general theory of relativity using which one obtains the following value for the limiting radius of the star:

$$R_s = \frac{2GM}{c^2} \quad (19)$$

which is known as the Schwarzschild radius. If the mass of the star is contained inside a sphere of radius

$$R < R_s$$

then a light beam will never leave the star and the star will be known as a black hole. Thus if

$$M = 10 M_\odot \approx 2 \times 10^{34} \text{ g}$$

where M_\odot ($\approx 2 \times 10^{33} \text{ g}$) represents the mass of the sun

$$R_s = \frac{2 \times 6.67 \times 10^{-8} \times 2 \times 10^{34}}{(3 \times 10^{10})^2} \text{ cm} \approx 30 \text{ km}$$

Indeed black holes with radius $\sim 10 \text{ km}$ have been detected!

*In the derivation of the Compton shift we have assumed that the electron is free although we know that the electrons are bound to the atoms. The assumption of a free electron is justified because the binding energy (\approx few eV) is usually very much smaller in comparison to the photon energy ($\geq 1000 \text{ eV}$).

SUMMARY

- In 1887, Herz discovered that a metal irradiated by a light beam would emit electrons. These electrons are known as photoelectrons and the phenomenon is known as *photoelectric effect*.
- There are certain peculiarities associated with the photoelectric effect which cannot be explained on the basis of wave theory. For example, a faint violet light would eject electrons of greater kinetic energy than an intense yellow light although the latter would produce a larger number of electrons. In 1905, Einstein provided a simple explanation of the peculiarities by assuming that light consisted of quanta of energy $h\nu$ (where ν is the frequency) and that the emission of a photoelectron was the result of the interaction of a single quantum (i.e. of the photon) with an electron. In his 1905 paper, Einstein wrote *Monochromatic radiation behaves as if it consists of mutually independent energy quanta of magnitude $[h\nu]$* .
- In 1923, Compton reported his studies on the scattering of X-rays by solid materials (mainly graphite) and showed that the shift of the wavelength of the scattered photon could be explained by assuming the photon to have momentum equal to h/λ . The Compton effect provided an unambiguous example of a process in which a quantum of radiation carrying energy as well as momentum scatters off an electron. The kinematics of the scattering process gives the following expression for the shift in the wavelength

$$\Delta\lambda = \frac{2h}{m_0 c} \sin^2 \frac{\theta}{2} \approx 0.0485 \sin^2 \frac{\theta}{2}$$

where θ is the angle of scattering of the light quantum, m_0 represents the rest mass of the electron and $\Delta\lambda$ is measured in Angstroms. Compton found the above formula to be in agreement with his experimental measurements of $\Delta\lambda$.

REFERENCES AND SUGGESTED READINGS

1. F.G. Smith and T.A. King, *Optics & Photonics*, John Wiley, Chichester, 2000.
2. W. H. Cropper, *The Quantum Physicists and an Introduction to Their Physics*, Oxford University Press, New York, 1970.
3. A. Einstein, 'On a heuristic point of view concerning the production and transformation of light', *Annalen der Physik*, Vol. 17, 132, 1905.
4. M. Jammer, *The Conceptual Development of Quantum Mechanics*, McGraw-Hill, New York, 1965.
5. R.A. Millikan, *The Electron and the light-quanta from the experimental point of view*, Nobel Lecture delivered in May 1924, Reprinted in *Nobel Lectures in Physics*, Elsevier Publishing Co., Amsterdam, 1965.
6. Bohm, *Quantum Theory*, Prentice-Hall, Englewood Cliffs, N.J., 1951.

PROBLEMS

- 22.1 (a) Calculate the number of photons emitted per second by a 5 mW laser assuming that it emits light of wavelength 6328 Å.
[Ans. 1.6×10^{16}]
- (b) The beam is allowed to fall normally on a plane mirror. Calculate the force acting on the mirror.
[Ans. 3.3×10^{-11} N]
- 22.2 Assume a 40 W sodium lamp ($\lambda \approx 5893 \text{ Å}$) emitting light in all directions. Calculate the rate at which the photons cross an unit area placed normally to the beam at a distance of 10 m from the source.
[Ans. $\approx 10^{17}$ photons/m²-sec]
- 22.3 In the photoelectric effect, a photon is completely absorbed by the electron. Show that the laws of conservation of energy and momentum cannot be satisfied simultaneously if a free electron is assumed to absorb the photon. (Thus the electron has to be bound to an atom and the atom undergoes a recoil when the electron is ejected. However, since the mass of the atom is much larger than that of the electron, the atom picks up only a small fraction of the energy, this is somewhat similar to the case of a tennis ball hitting a heavy object, the momentum of the ball is reversed with its energy remaining almost the same.)
- 22.4 In the Compton scattering experiment, show that the fractional loss of energy of the photon increases with decrease in the wavelength. Calculate the maximum value of this fractional loss for $\lambda = 0.711 \text{ Å}$ and $\lambda = 0.022 \text{ Å}$; the former corresponds to the Molybdenum K_α X-ray line and the latter to the γ -rays emitted from RaC.
- 22.5 If photoelectrons are emitted from a metal surface by using blue light, can you say for sure that photoelectric emission will take place with yellow light and with violet light?

7. R.S. Shankland (Editor), *Scientific Papers of A. H. Compton: X-ray and Other Studies*, University of Chicago Press, Chicago; 1975. [Most of the papers of A.H. Compton are reprinted here].
8. M. Born, *Atomic Physics*, Blackie & Son, London, 1962.
9. A.H. Compton, *A Quantum theory on the scattering of X-rays by light elements*, Physical Review, Vol. **21**, 483, 1923, Reprinted in Ref. 7.
10. A.H. Compton, *The Spectrum of Scattered X-rays*, Physical Review, Vol. **22**, 409, 1923. Reprinted in Ref. 7.



PART 7 **Sources of** **Coherent Light**

This part consists of only one chapter on Lasers – the discovery of which in 1960 has created a revolution in many diverse areas. The subject is of such great technological importance that Townes, Prochorov and Basov shared the 1964 Nobel Prize in Physics for developing lasers. The chapter discusses the basic physics of lasers along with their special characteristics and numerous applications.

Chapter 23

Lasers: An Introduction

In The War of Worlds, written before the turn of the century, H.G. Wells told a fanciful story of how Martians invaded and almost conquered the earth. Their weapon was a mysterious 'sword of heat', from which flickered 'a ghost of a beam of light', it felled men in their tracks, made lead run like water and flashed anything combustible into masses of flame. Today Wells' sword of heat comes close to reality in the laser...

Thomas Meloy

Important Milestones

- 1917 The theory of stimulated emission was put forward by Albert Einstein.
- 1954 The phenomenon of stimulated emission was first used by Charles Townes in 1954 in the construction of a microwave amplifier device called the maser which is an acronym for **M**icrowave **A**mplification by **S**timulated **E**mission of **R**adiation. At about the same time, a similar device was also proposed by Prochorov and Basov in USSR.
- 1958 The maser principle was later extended to the optical frequencies by Schawlow and Townes in 1958, which led to the realization of the device now known as the laser. Townes, Basov and Prochorov were awarded the 1964 Nobel Prize in physics for their *fundamental work in the field of Quantum Electronics, which has led to the construction of oscillators and amplifiers based on the laser-maser principle.**
- 1960 The first successful operation of a laser device ($\lambda \sim 0.684 \mu\text{m}$) was demonstrated by Theodore Maiman in 1960 using a ruby crystal (see Sec. 23.3).
- 1961 Within a few months of the operation of the ruby laser, Ali Javan and his associates constructed the first gas laser ($\lambda \sim 0.633 \mu\text{m}$), namely the helium-neon laser (see Sec. 23.4).
- 1961 The first fiber laser (barium crown glass doped with Nd^{3+} ions) was fabricated by Elias Snitzer (see Sec. 23.2).
- 1962 Semiconductor laser (which are now extensively used in fiber-optic communication systems) was discovered by four independent groups.
- 1963 C.K.N. Patel discovered the CO_2 laser ($\lambda \sim 10.6 \mu\text{m}$).
- 1964 W. Bridges discovered the Ar-ion laser ($\lambda \sim 0.515 \mu\text{m}$);
J.E. Geusic and his co-workers discovered the Nd:YAG laser ($\lambda \sim 0.515 \mu\text{m}$).
Since then, laser action has been obtained in a large variety of materials including liquids, ionized gases, dyes, semiconductors, etc.

23.1 INTRODUCTION

LASER is an acronym for **L**ight **A**mplification by **S**timulated **E**mission of **R**adiation. The light emitted from a laser often possesses some very special characteristics—some of these are:

(a) **Directionality:** The divergence of the laser beam is usually limited by diffraction (see Sec. 16.4) and the

actual divergence can be less than 10^{-5} radians; this leads to the application of the laser in surveying, remote sensing, lidar, etc.

(b) **High Power:** Continuous wave lasers having power levels $\sim 10^5 \text{ W}$ and pulsed lasers having a total energy $\sim 50,000 \text{ J}$ can have applications in welding, cutting, laser fusion, star wars etc.

(c) **Tight Focusing:** Because of highly directional properties of the laser beams, they can be focused to areas

*The Nobel lectures of Townes, Basov and Prochorov [1-3] give a nice perspective of the field; these are reprinted in Ref. 4.

$\sim \text{few } (\mu\text{m})^2$ – this leads to applications in surgery, material processing, compact discs, etc.

- (d) **Spectral Purity:** Laser beams can have an extremely small spectral width $\Delta\lambda \sim 10^{-6} \text{ \AA}$. Because of high spectral purity, lasers find applications in holography, optical communications, spectroscopy etc.

Because of such unique properties of the laser beam, it finds important applications in many diverse areas and indeed one can say that after the discovery of the laser, optics has become an extremely important field of study. For example, in Example 16.5 we had shown that a 2 mW diffraction limited laser beam incident on the eye can produce an intensity of about 10^6 W/m^2 at the retina—this would certainly damage the retina. Thus, whereas it is quite safe to look at a 500 W bulb, it is very dangerous to look directly into a 5mW laser beam. Indeed, because a laser beam can be focused to very narrow areas, it has found applications in areas like eye surgery, laser cutting, etc.

The basic principle involved in the lasing action is the phenomenon of stimulated emission, which was predicted by Einstein in 1917.*⁵ In this introductory section we will discuss this first, which will be followed by brief discussions of the main components of a laser and the underlying principle as to how the laser works. In Sec 23.2 we will briefly discuss the working of a fiber laser and in Sec 23.3 we will discuss the working of the ruby laser, which was the first laser to be fabricated. In Sec 23.4 we will discuss the working of the helium-neon laser. In Sec. 23.5 we will have a slightly more detailed account of resonators and in Sec. 23.6 we will discuss Einstein coefficients and optical amplification. In Sec. 23.7 we will discuss the line shape function and finally in Sec. 23.8 we will discuss the monochromaticity of the laser beam.

23.1.1 Spontaneous and Stimulated Emissions

Atoms are characterised by discrete energy states. According to Einstein, there are three different ways in which an atom can interact with electromagnetic radiation:

(a) Spontaneous Emission

Atoms in the energy state E_2 can make a (spontaneous) transition to the energy state E_1 with the emission of radiation of frequency

$$\omega = \frac{E_2 - E_1}{\hbar} \quad (1)$$

where

$$\hbar = \frac{h}{2\pi} \approx 1.0546 \times 10^{-34} \text{ Js}$$

*The original paper of Einstein is reprinted in Ref. 6.

and $h (= 6.626 \times 10^{-34} \text{ Js})$ is known as the Planck's constant. Since this process can occur even in the absence of any radiation, this is called spontaneous emission [see Fig. 23.1(a)]. The rate of spontaneous emission is proportional to the number of atoms in the excited state.

(b) Stimulated Emission

As put forward by Einstein, when an atom is in the excited state, it can also make a transition to a lower energy state through what is known as stimulated emission in which an incident signal of appropriate frequency triggers an atom in an excited state to emit radiation—this results in the amplification of the incident beam [see Fig. 23.1(b)]. The rate of stimulated emission depends both on the intensity of the external field and also on the number of atoms in the excited state.

(c) Stimulated Absorption

Stimulated absorption (or simply absorption) is the process in which the electromagnetic radiation of an appropriate frequency (corresponding to the energy difference of the two atomic levels) can pump the atom to its excited state [see Fig. 23.1(c)]. The rate of stimulated absorption depends both on the intensity of the external field and also on the number of atoms in the lower energy state.

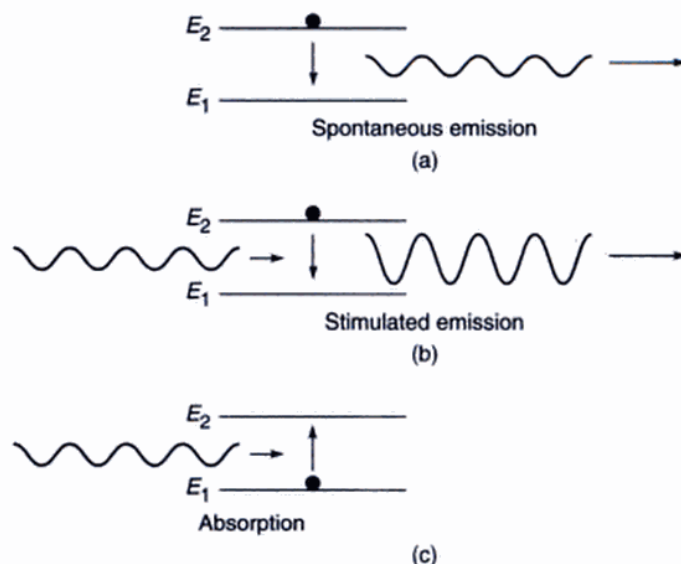


Fig. 23.1 (a) Spontaneous emission, (b) stimulated emission and (c) stimulated absorption.

When the atoms are in thermodynamic equilibrium, there are larger number of atoms in the lower state implying that the number of absorptions exceeds the number of stimulated emissions; this results in the attenuation of the beam

[see Fig. 23.2(a)]. On the other hand, if we are able to create a state of population inversion in which there are larger number of atoms in the upper state then the number of stimulated emissions would exceed the number of absorptions resulting in the (optical) amplification of the beam [see Fig. 23.2(b)]. The amplification process due to stimulated transitions is *phase coherent*, i.e., the energy delivered by the molecular system has the same field distribution and frequency as the stimulating radiation.¹

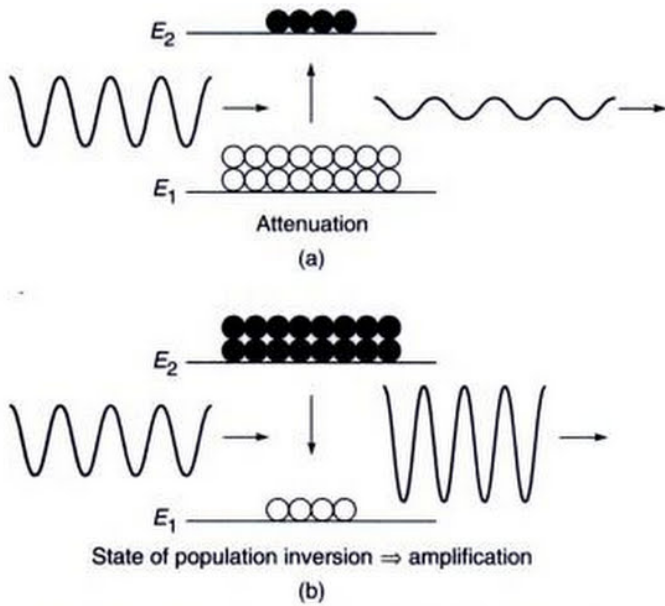


Fig. 23.2 (a) Larger number of atoms in the lower state result in the attenuation of the beam. (b) Larger number of atoms in the upper state (which is known as population inversion) result in the amplification of the beam.

23.1.2 Main Components of the Laser

The three main components of any laser are (see Fig. 23.3):

1. **The active medium:** The active medium consists of a collection of atoms, molecules or ions (in solid, liquid or gaseous form), which is capable of amplifying light waves. Under normal circumstances, there are always a larger number of atoms in the lower energy state than in the excited energy state. An electromagnetic wave passing through such a collection of atoms would get attenuated; this is discussed in detail in Sec. 23.6. In order to have optical amplification, the

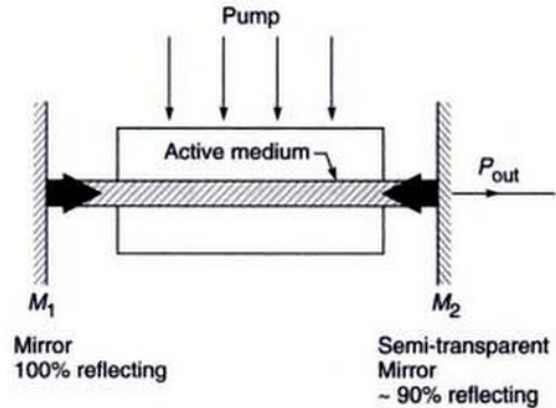


Fig. 23.3 The three basic components of a laser are (i) the active medium (which provides amplification), (ii) the optical resonator (which provides frequency selection and optical feedback) and (iii) the pump (which supplies power to the active medium to achieve population inversion).

medium has to be kept in a state of *population inversion*, i.e., in a state in which the number of atoms in the upper energy level is greater than that in the lower energy level—this is achieved by means of the pump.

2. **The pumping source:** The pumping mechanism provides for obtaining such a state of population inversion between a pair of energy levels of the atomic system and when we have a state of population inversion, the input light beam can get amplified by stimulated emission (see Fig. 23.4).

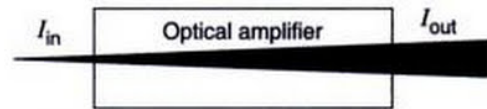


Fig. 23.4 The active medium essentially consists of a collection of atoms in a state of population inversion which can amplify the input light beam (or spontaneously emitted light) by stimulated emission. This is known as optical amplification.

3. **The optical resonator:** A medium with population inversion is capable of amplification; however, in order that it acts as an oscillator, a part of the output energy must be fed back into the system.* Such a

* Since some of the energy is coupled back to the system, it is said to act as an oscillator. Indeed, in the early stages of the development of the laser, there was a move to change its name to LOSER which is an acronym for Light Oscillation by Stimulated Emission of Radiation. Since it would have been difficult to obtain a research grant for LOSERS, it was decided to retain the name LASER.

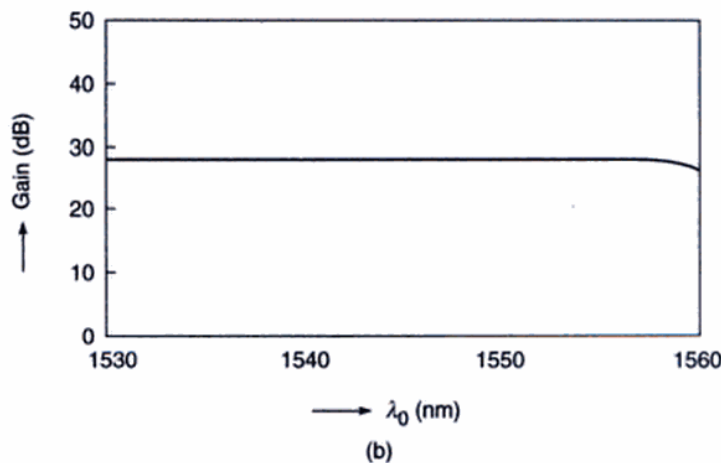
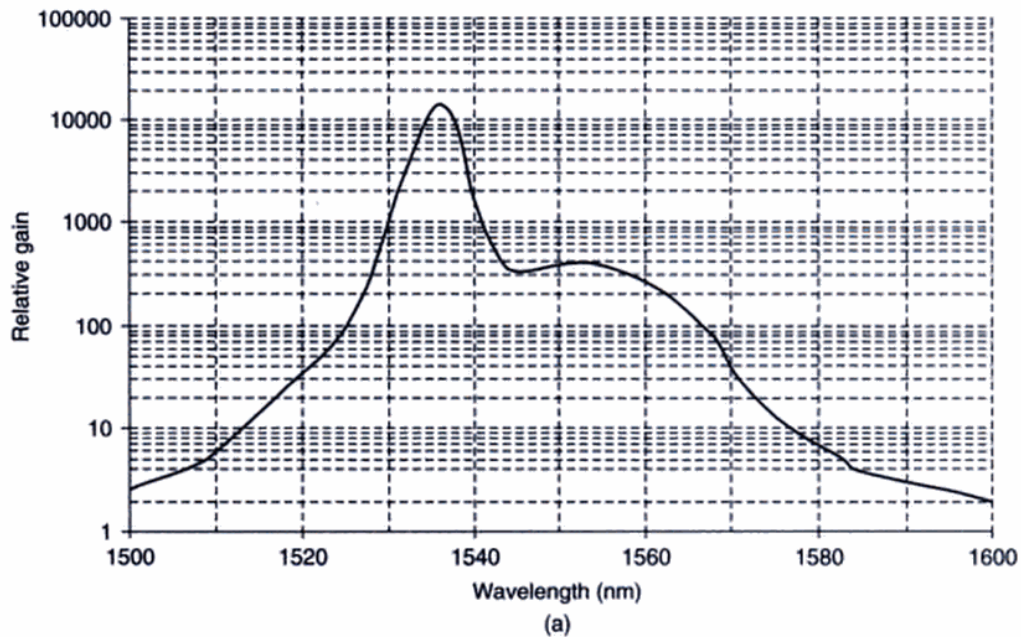


Fig 23.8 (a) The gain spectrum of a typical erbium doped fiber amplifier using a 50 mW pump at 980 nm (Adapted from Ref. 10). (b) Through various mechanisms, the gain spectrum of an EDFA can be made almost flat. The above figure corresponds to an EDFA which has an almost flat gain (of about 28 dB) in the wavelength region 1530 nm to 1560 nm (Adapted from Ref. 11).

amplification coefficient due to the gain and a certain attenuation coefficient due to the losses in the cavity. The modes for which the losses in the cavity exceed the gain die out. On the other hand, the modes whose gain is higher than the losses get amplified by drawing energy from the laser medium. The amplitude of the mode increases rapidly until the upper level population reaches a value when the gain equals the losses, and the mode oscillates in steady state. When the laser oscillates in the steady state, the losses are exactly compensated for by the gain provided by the medium, and the wave coming out of the laser can be represented as a continuous wave.

23.2 THE FIBER LASER

If we put the doped fiber between two mirrors (which act as a resonator)—then with an appropriate pump we would have a fiber laser (see Fig. 23.9). Indeed in 1961, Elias Snitzer wrapped a flashlamp around a glass fiber (having a $300\text{ }\mu\text{m}$ core doped with Nd^{3+} ions clad in a lower index glass) and when suitable feedback was applied, the first fiber laser was born.¹² Thus the fiber laser was fabricated within a year of the demonstration of the first ever laser by Theodore Maiman. These days fiber lasers are commercially available in the market which have applications in

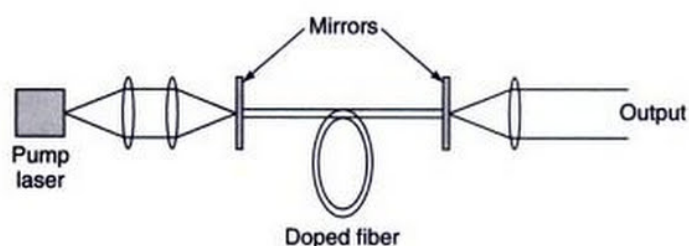


Fig 23.9 A fiber laser—the output power can be as high as 10 W.

many diverse areas because of their flexibility and high power levels. The lower curve in Fig. 23.10 corresponds to the output spectrum of an EDFA just before it starts lasing. As we increase the pump power, the EDFA starts lasing and the spikes correspond to the various resonator modes; the ends of the fiber act as the resonator.

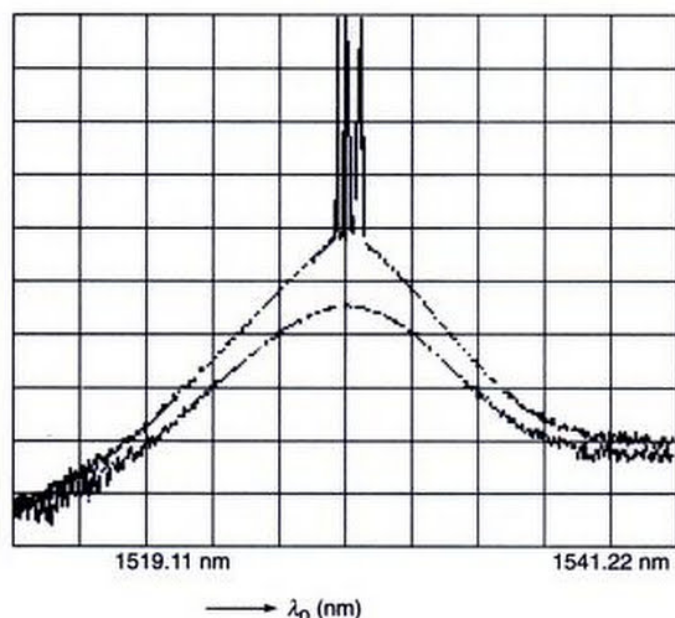


Fig. 23.10 The lower and upper curves show the output of an EDFA just before and after it starts lasing. (Photograph courtesy Professor Thyagarajan and Mr. Mandeep Singh).

23.3 THE RUBY LASER

In the first laser fabricated by Maiman in 1960¹³, the population inversion was achieved in the following manner.

*The Al_2O_3 crystal which serves as a medium to suspend the chromium ions is known as the host crystal. The characteristics of the host crystal affect the laser action and also the broadening of the energy levels of the activator atoms which in this case is chromium. For a good lasing action, the ruby crystal consists of about 0.05% (by weight) of chromium; however, higher concentrations of chromium have also been used. For a detailed discussions of host crystals, see Ref. 14.

It was made from a single cylindrical crystal of ruby whose ends were flat, with one of the ends completely silvered and the other partially silvered (see Fig. 23.11). Ruby consists of Al_2O_3 with some of the aluminum atoms replaced by chromium.* The energy states of the chromium ion are shown in Fig. 23.12. The chief characteristic of the energy levels of a chromium ion is the fact that the bands labeled E_1 and E_2 have a lifetime of $\sim 10^{-8}$ sec whereas the state marked M has a lifetime of $\sim 3 \times 10^{-3}$ sec—the lifetime represents the average time an atom spends in an excited state before making a transition to a lower energy state. A state characterized by such a long lifetime is termed a metastable state.

The chromium ion in its ground state can absorb a photon (whose wavelength is around 6600 Å) and make a

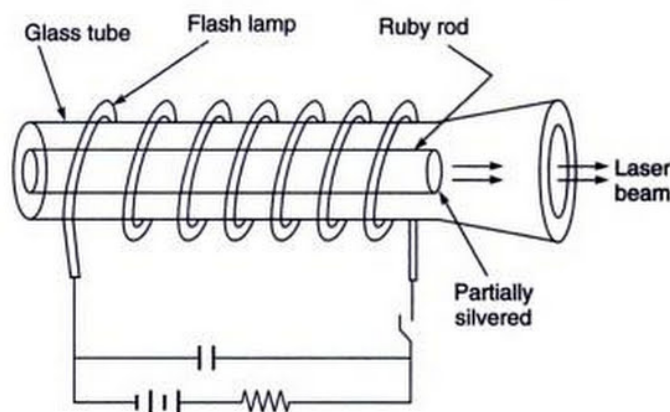


Fig. 23.11 The ruby laser

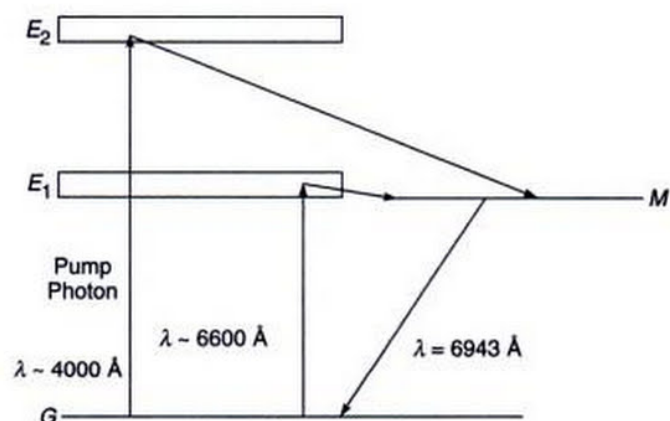


Fig. 23.12 The energy levels of the chromium ion; G and M represent the ground and meta-stable states respectively.

transition to one of the states in the band E_1 ; it could also absorb a photon of $\lambda \sim 4000 \text{ \AA}$ and make a transition to one of the states in the band E_2 —this is known as optical pumping and the photons which are absorbed by the chromium ions are produced by the flash lamp (see Fig. 23.11). In either case, it immediately makes a non-radiative transition (in a time $\sim 10^{-8}$ sec) to the metastable state M —in a non-radiative transition, the excess energy is absorbed by the lattice and does not appear in the form of electromagnetic radiation. Also since the state M has a very long life, the number of atoms in this state keeps increasing and one may achieve population inversion between states M and G . Thus we may have a larger number of atoms in states M and G . Once population inversion is achieved, light amplification can take place, with two reflecting ends of the ruby rod forming a cavity. The ruby laser is an example of a three level laser.

In the original setup of Maiman, the flashlamp (filled with xenon gas) was connected to a capacitor (see Fig. 23.11) which was charged to a few kilovolts. The energy stored in the capacitor (\sim a few thousand joules) was discharged through the xenon lamp in a few milliseconds. This results in a power which is \sim a few megawatts. Some of this energy is absorbed by the chromium ions resulting in their excitation and subsequent lasing action.

23.3.1 Spiking in Ruby Laser

The flash operation of the lamp leads to a pulsed output of the laser. Even in the short period of a few tens of microseconds in which the ruby is lasing, one finds that the emission is made up of spikes of high intensity emissions as shown in Fig. 23.13. This phenomenon is known as spiking and can be understood as follows. When the pump is suddenly switched on to a value much above the threshold, the population inversion builds up and crosses the threshold value,

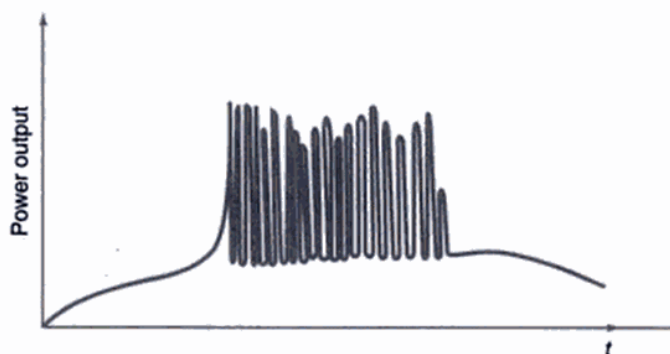


Fig. 23.13 The characteristic spiking of a ruby laser.

as a consequence of which the photon number builds up rapidly to a value much higher than the steady state value. Since the photon number is higher than the steady state value, the rate at which the upper level depletes (because of stimulated transitions) is much higher than the pump rate. Consequently, the inversion becomes below threshold and the laser action ceases. Thus the emission stops for a few microseconds, within which time the flashlamp again pumps the ground state atoms to the upper level, and laser oscillations begin again. This process repeats itself till the flashlamp power falls below the threshold value and the lasing action stops (see Fig. 23.13).

23.4 THE He-Ne LASER

We will now briefly discuss the He-Ne laser which was first fabricated by Ali Javan and his coworkers at Bell Telephone Laboratories in USA.¹⁵ This was also the first gas laser to be operated successfully.

The He-Ne laser consists of a mixture of He and Ne in a ratio of about 10:1, placed inside a long narrow discharge tube (see Fig. 23.14). The pressure inside the tube is about 1 Torr.* The gas system is enclosed between a pair of plane mirrors or a pair of concave mirrors so that a resonator system is formed. One of the mirrors is of very high reflectivity while the other is partially transparent so that energy may be coupled out of the system.

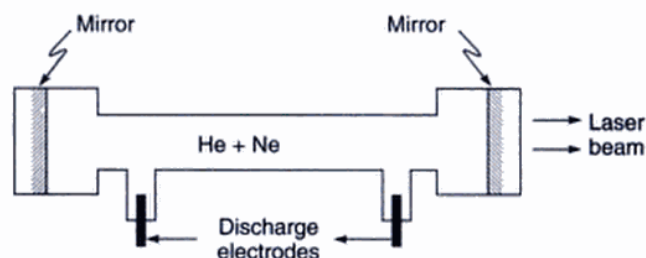


Fig. 23.14 The helium-neon laser.

The first few energy levels of He and Ne atoms are shown in Fig. 23.15. When an electric discharge is passed through the gas, the electrons traveling down the tube collide with the He atoms and excite them (from the ground state F_1) to the levels marked F_2 and F_3 . These levels are metastable, i.e., He atoms excited to these states stay in these levels for a sufficiently long time before losing energy through collisions. Through these collisions, the Ne atoms are excited to the levels marked E_4 and E_6 which have nearly the same energy as the levels F_2 and F_3 of He. Thus

* 1 Torr = 1 mm of Hg = 133 Pascal = 133 N/m²; the unit 'Torr' is named after Torricelli, the seventh century Italian mathematician who invented the mercury manometer.

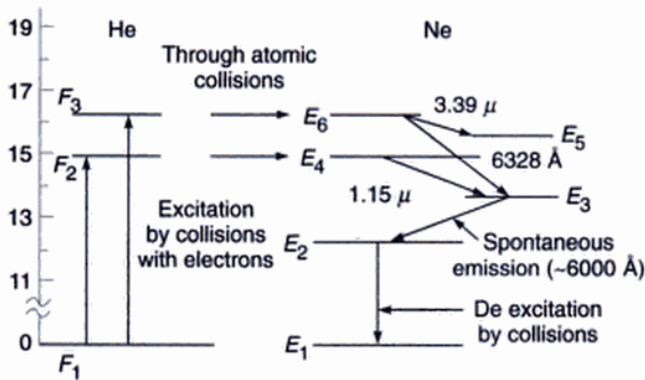


Fig. 23.15 Relevant energy levels of helium and neon.

when the atoms in levels F_2 and F_3 collide with unexcited Ne atoms, they raise them to the levels E_4 and E_6 respectively. Thus, we have the following two step process:

1. Helium atom in the ground state F_1 + collision with electron
 \rightarrow Helium atom in the excited state (F_2 or F_3)
 + electron with lesser kinetic energy.
2. The excited states of He (F_2 or F_3) are metastable*—they would not readily lose energy through spontaneous emissions (the radioactive life time of these excited states would be about one hour). However, they can readily lose energy through collisions with Ne atoms:

He atom in the excited state F_3 + Ne atom in the ground state

\rightarrow He atom in the ground state
 + Ne atom in the excited state E_6 .

Similarly

He atom in the excited state F_2 + Ne atom in the ground state

\rightarrow He atom in the ground state
 + Ne atom in the excited state E_4 .

This results in a sizeable population of the levels E_4 and E_6 . The population in these levels happens to be much more than those in the lower levels E_3 and E_5 . Thus a state of population inversion is achieved and any spontaneously emitted photon can trigger laser action in any of the three transitions shown in Fig. 23.15. The Ne atoms then drop down from the lower laser levels to the level E_2 through spontaneous emission. From the level E_2 the Ne atoms are brought back to the ground state through collision with the

walls. The transition from E_6 to E_5 , E_4 to E_3 and E_6 to E_3 result in the emission of radiation having wavelengths $3.39 \mu\text{m}$, $1.15 \mu\text{m}$ and 6328 \AA respectively. It may be noted that the laser transitions corresponding to $3.39 \mu\text{m}$ and $1.15 \mu\text{m}$ are not in the visible region. The 6328 \AA transition corresponds to the well-known red light of the He-Ne laser. A proper selection of different frequencies may be made by choosing end mirrors having high reflectivity over only the required wavelength range. The pressures of the two gases must be so chosen that the condition of population inversion is not quenched. Thus the conditions must be such that there is an efficient transfer of energy from He to Ne atoms. Also, since the level marked E_2 is metastable, electrons colliding with atoms in level E_2 may excite them to level E_3 , thus decreasing the population inversion. The tube containing the gaseous mixture is also made narrow so that He atoms in level E_2 can get de-excited by collision with the walls of the tube. Referring to Fig. 23.15, it may be mentioned that actually there are a large number of levels grouped around E_2 , E_3 , E_4 , E_5 and E_6 . Only those levels are shown in the figure which correspond to the important laser transitions.**

Gas lasers are, in general, found to emit light, which is more directional and more monochromatic. This is because of the absence of such effects as crystalline imperfection, thermal distortion and scattering, which are present in solid-state lasers. Gas lasers are capable of operating continuously without need for cooling.

23.5 OPTICAL RESONATORS

In Sec. 23.1 we had briefly discussed that a light beam passing through a suitable medium with population inversion may be amplified. In order to construct an oscillator, which can supply light energy and act as a source of light, one must couple a part of the output back into the medium. This can be achieved by placing the active medium between two mirrors which reflect most of the **output energy** back to the system—see Fig. 23.3. Such a system of two mirrors represents a *resonant cavity*.

Now, in order to obtain an output beam, one of the mirrors is made partially reflecting. Thus, imagine a wave that starts from one of the mirrors and travels towards the other. In passing through the active medium, it gets amplified. If the second mirror is partially reflecting, then the wave is partially transmitted and the rest reflected back towards the first mirror. In traveling to the first mirror, it again gets amplified and returns to the position it has started from. Thus, in between the two mirrors, we have waves

*The spectroscopic states corresponding to the states F_1 , F_2 and F_3 are 1^1S_0 , 2^3S_1 and 2^1S_0 respectively.

**Further details on He-Ne Laser can be found in Refs 16 and 17.

or

$$u(\omega) = \frac{A_{21}}{\frac{N_1}{N_2} B_{12} - B_{21}} \quad (22)$$

Now according to a fundamental principle in thermodynamics, at thermal equilibrium, we have the following expression for the ratio of the populations of two levels:

$$\frac{N_1}{N_2} = \exp\left[\frac{E_2 - E_1}{k_B T}\right] = \exp\left[\frac{\hbar\omega}{k_B T}\right] \quad (23)$$

where $k_B (= 1.38 \times 10^{-23} \text{ J/K})$ represents the Boltzmann constant and T represents the absolute temperature. Equation (21) is known as the Boltzmann's law. Thus, we may write

$$u(\omega) = \frac{A_{21}}{B_{12} e^{\hbar\omega/(k_B T)} - B_{21}} \quad (24)$$

Now, at thermal equilibrium, the radiation energy density is given by Planck's law:

$$u(\omega) = \frac{\hbar\omega^3 n_0^3}{\pi^2 c^3} \frac{1}{e^{\hbar\omega/(k_B T)} - 1} \quad (25)$$

where n_0 represents the refractive index of the medium. Comparing Eqs. (24) and (25), we obtain*

$$B_{12} = B_{21} = B \quad (\text{say}) \quad (26)$$

and

$$\frac{A_{21}}{B_{21}} = \frac{\hbar\omega^3 n_0^3}{\pi^2 c^3} \quad (27)$$

Notice that if we had not assumed the presence of stimulated emission we would not have been able to arrive at an expression for $u(\omega)$; Einstein in 1917 had predicted the existence of stimulated emission which was later confirmed by rigorous quantum theory. (See Chapter 27 of Ref. 19).

It may be noted that at thermal equilibrium the ratio of the number of spontaneous to stimulated emissions is given by

$$\frac{A_{21}}{B_{21} u(\omega)} = e^{\hbar\omega/(k_B T)} - 1 \quad (28)$$

We may note the following two important points:

1. For normal optical source, $T \sim 10^3 \text{ }^\circ\text{K}$ with $\omega \approx 3 \times 10^{15} \text{ sec}^{-1}$ (corresponding to $\lambda \approx 6000 \text{ \AA}$) we have

$$\frac{\hbar\omega}{k_B T} \approx \frac{1.054 \times 10^{-34} (\text{J} \cdot \text{sec}) \times 3 \times 10^{15} \text{ sec}^{-1}}{1.38 \times 10^{-23} (\text{J}/^\circ\text{K}) \times 10^3} \approx 23$$

giving

$$\frac{A_{21}}{B_{21} u(\omega)} = 10^{10}$$

Thus, when the atoms are in thermal equilibrium, the emission (at optical frequencies) is predominantly due to spontaneous transitions and hence the emission from ordinary light sources is incoherent.

2. From Eq. (27), one can see that the coefficient B_{21} is inversely proportional to ω^3 implying that laser action would become more difficult as we go to higher frequencies.

23.6.1 Population Inversion

In the previous section we had assumed that the atom is capable of interacting with radiation of a particular frequency ω . However, if one observes the spectrum of the radiation due to spontaneous emissions from a collection of atoms, one finds that the radiation is not monochromatic but is spread over a certain frequency range. This would imply that energy levels have widths and atoms can interact over a range of frequencies. As an example, in Fig. 23.22 we have shown that the $2P$ level of hydrogen atom has a certain width $\Delta E (= \hbar\Delta\omega)$ so that the atom can absorb/emit radiation over a range of frequencies $\Delta\omega$. For the $2P \rightarrow 1S$ transition

$$\Delta E = 4 \times 10^{-7} \text{ eV} \Rightarrow \Delta\omega \approx 6 \times 10^8 \text{ s}^{-1}$$

Since $\omega_0 \approx 1.55 \times 10^{16} \text{ s}^{-1}$, we get

$$\frac{\Delta\omega}{\omega_0} \approx 4 \times 10^{-8}$$

Thus, in general, $\Delta\omega \ll \omega_0$ showing the spectral purity of the source. We introduce the normalized line shape function $g(\omega)$ such that

Number of spontaneous emissions/unit time/unit volume so that the emitted frequency lies between ω and $\omega + d\omega$

$$= N_2 A_{21} g(\omega) d\omega$$

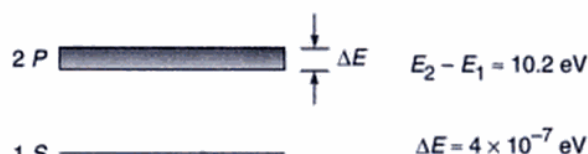


Fig. 23.22 The $2P$ level of hydrogen atom has a certain width $\Delta E (= \hbar\Delta\omega)$ so that the atom can absorb/emit radiation over a range of frequencies $\Delta\omega$.

*If the levels 1 and 2 are g_1 and g_2 fold degenerate, then $N_1/N_2 = (g_1/g_2) \exp(\hbar\omega/k_B T)$, $B_{12} = B_{21} (g_2/g_1)$ and $A_{21}/B_{21} = n_0^3 \hbar\omega^3/\pi^2 c^3$.

Similarly,

Number of stimulated emissions/unit time/unit volume so that the emitted frequency lies between ω and $\omega + d\omega$

$$= N_2 B_{21} u(\omega) g(\omega) d\omega$$

Number of stimulated absorptions/unit time/unit volume so that the absorbed frequency lies between ω and $\omega + d\omega$

$$= N_1 B_{12} u(\omega) g(\omega) d\omega$$

Thus the total number of stimulated emissions/unit time/unit volume will be given by

$$\begin{aligned} W_{21} &= N_2 \int_0^\infty B_{21} u(\omega) g(\omega) d\omega \\ &= N_2 \frac{\pi^2 c^3}{\hbar t_{sp} n_0^3} \int_0^\infty \frac{u(\omega)}{\omega^3} g(\omega) d\omega \end{aligned}$$

where we have used Eqs (27) and (20). Now, for a near monochromatic radiation field (as it is indeed the case for the laser), $u(\omega)$ is very sharply peaked at a particular value of ω (say ω') and in carrying out the above integration, $g(\omega)/\omega^3$ can be assumed to be essentially constant over the region where $u(\omega)$ is appreciable to give

$$W_{21} \approx N_2 \frac{\pi^2 c^3}{\hbar t_{sp} n_0^3} \frac{g(\omega')}{\omega'^3} U \quad (29)$$

where $g(\omega')$ represents the value of the line-shape function evaluated at the radiation frequency ω' and U represents the energy density associated with the radiation field.*

$$U = \int_0^\infty u(\omega) d\omega \quad (30)$$

Now the energy density U and the intensity I_ω are related through the following equation** [see Eq. (78) of Chapter 20]

$$I_\omega = v U = \frac{c}{n_0} U \quad (31)$$

where $v (= c/n_0)$ represents the velocity of the radiation field in the medium, n_0 being its refractive index. [The quantity I_ω represents energy per unit area per unit time, the MKS units of I_ω would therefore be $\text{Jm}^{-2}\text{s}^{-1}$; we may

*The argument essentially implies

$$u(\omega) = U \delta(\omega - \omega')$$

where $\delta(\omega - \omega')$ represents the Dirac-delta function.

**This is analogous to the equation $J = \rho v$, where ρ represents the number of particles per unit volume (all propagating with velocity v) and J represents the number of particles crossing a unit area perpendicular to the direction of propagation per unit time. This can be easily seen from the fact that the number of particles crossing a unit area per unit time would be those contained in a cylinder of length v units with unit area of cross-section.

mention that the quantity U is denoted by $\langle u \rangle$ in Sec. 20.5]. Thus the total number of stimulated emissions/unit time/unit volume will be given by

$$W_{21} = N_2 \frac{\pi^2 c^2}{\hbar t_{sp} n_0^2} \frac{g(\omega)}{\omega^3} I_\omega \quad (32)$$

where we have dropped the prime on ω . Similarly, the number of stimulated absorptions per unit time per unit volume would be given by

$$W_{12} = N_1 \frac{\pi^2 c^2}{\hbar t_{sp} n_0^2} \frac{g(\omega)}{\omega^3} I_\omega \quad (33)$$

We next consider a collection of atoms and let a near monochromatic beam of frequency ω be propagating through it along the z -direction. In order to obtain an expression for the rate of change of the intensity of the beam as it propagates, we consider two planes of area S perpendicular to the z -direction at z and $z + dz$ (see Fig. 23.23). The volume of the medium between planes P_1 and P_2 is $S dz$ and hence the number of stimulated absorptions per unit time is $W_{12} S dz$. Since each photon has an energy $\hbar\omega$, the energy absorbed per unit time in the volume element $S dz$ is

$$W_{12} \hbar\omega S dz$$

Similarly the corresponding energy gain (because of stimulated emissions) is

$$W_{21} \hbar\omega S dz$$

where we have neglected the radiation arising out of spontaneous emissions, because such radiations propagate

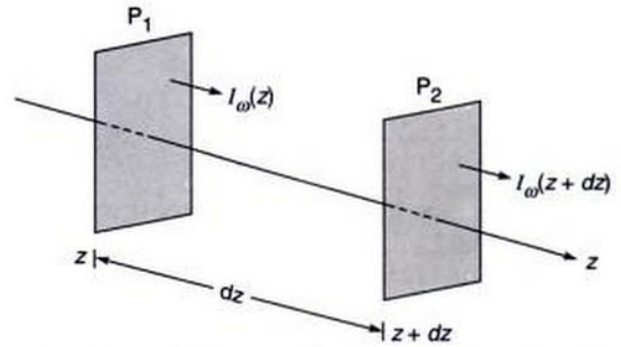


Fig. 23.23 Electromagnetic wave propagating along the z -direction through a collection of atoms.

in random directions and are, in general, lost from the beam. Thus, the net amount of energy absorbed per unit time in the volume element $S dz$ is

$$(W_{12} - W_{21}) \hbar \omega S dz$$

If $I_\omega(z)$ represents the intensity of the beam in the plane P_1 , then the total energy entering the volume element $S dz$ per unit time is

$$I_\omega(z)S$$

Similarly, if $I_\omega(z + dz)$ represents the intensity in the plane P_2 , then the total energy leaving the volume element per unit time is

$$I_\omega(z + dz)S = I_\omega(z)S + \frac{\partial I_\omega}{\partial z} dz S$$

Hence the net amount of energy leaving the volume element per unit time is

$$\frac{\partial I_\omega}{\partial z} dz S$$

This must be equal to the negative of the energy absorbed by the medium between z and $z + dz$. Thus,

$$\begin{aligned} \frac{\partial I_\omega}{\partial z} S dz &= -(W_{12} - W_{21}) \hbar \omega S dz \\ &= -\frac{\pi^2 c^2}{\hbar t_{sp} \omega^3 n_0^2} g(\omega) I_\omega \hbar \omega S dz (N_1 - N_2) \end{aligned}$$

or

$$\frac{I}{I_\omega} \frac{\partial I_\omega}{\partial z} = \gamma \quad (34)$$

where

$$\gamma = \frac{\pi^2 c^2}{\omega^2 t_{sp} n_0^2} (N_2 - N_1) g(\omega) \quad (35)$$

Since the line shape function $g(\omega)$ is very sharply peaked (see Section 23.6.2), the function γ is also a sharply peaked function. Equation (34) can be readily integrated to give*

$$I_\omega(z) = I_\omega(0) e^{\gamma z} \quad (36)$$

Thus if $N_1 > N_2$, γ is negative and the intensity of the beam decreases exponentially with z ; the intensity decreasing to $1/e$ of its value at $z = 0$ in a distance $1/\gamma$. Hence at thermal equilibrium, since the number of atoms in the lower level is greater than that in the upper level, the intensity of the beam (as it propagates through the medium) decreases

exponentially. On the other hand, if there are more atoms in the higher energy level than in the lower level, (i.e., if there is a population inversion) then $\gamma > 0$ and there would be an exponential increase in intensity of the beam; this is known as light amplification.

23.6.2 Cavity Lifetime

In an actual laser system, the active medium (which is capable of amplification) is placed between a pair of mirrors forming what is known as a resonator (see Sec. 23.5). In order that oscillations be sustained in the cavity, it is essential that the net losses suffered by the beam be compensated for by the gain of the medium. At threshold and under steady state operation, the two are exactly compensated. In order to obtain the threshold condition, we first calculate the passive cavity lifetime t_c which is the time in which the energy $W(t)$ in the (passive) cavity decreases by a factor $1/e$ in the absence of the amplifying medium:

$$W(t) = W(0) \exp[-t/t_c] \quad (37)$$

Let d represent the length of the active medium. In one round trip the beam traverses a distance $2d$ through the active medium and gets attenuated by a factor

$$R_1 R_2 \exp[-2\alpha_c d]$$

where R_1 and R_2 are the reflectivities of the mirrors at the two ends of the resonator and the term $\exp[-2\alpha_c d]$ represents losses caused by absorption, scattering, diffraction, etc. Now the time taken for one round trip is given by

$$t = \frac{2d}{c/n_0}$$

Thus

$$\exp\left[-\frac{2d}{(c/n_0)t_c}\right] = R_1 R_2 \exp[-2\alpha_c d] \quad (38)$$

giving the following expression for the passive cavity lifetime:

$$\frac{1}{t_c} = \frac{c/n_0}{2d} [2\alpha_c d - \ln(R_1 R_2)] \quad (39)$$

It can be easily seen that the cavity lifetime can also be expressed as

$$t_c = \frac{2n_0 d}{c \ln\left(\frac{1}{1-x}\right)} \quad (40)$$

*In obtaining Eq. (36) from Eq. (34), it has been assumed that $N_1 - N_2$ (and hence γ) is independent of I_ω . Such an approximation is valid only for small values of I_ω . For intense light beams (when I_ω becomes very large) saturation of the levels set in and the attenuation is linear rather than exponential (see, for example, Ref. 4, Sections 4.2 and 4.3).

Example 23.3 Typical parameters for a He-Ne laser

We consider a He-Ne laser; we assume $T \approx 300^\circ \text{K}$. Thus, for the $\lambda_0 \approx 6328 \text{ \AA}$ radiation

$$\begin{aligned}\Delta\omega_D &= \frac{2\omega_{21}}{c} \left[\frac{2k_B T}{M} \ln 2 \right]^{\frac{1}{2}} \\ &= \frac{4\pi}{\lambda_0} \left[\frac{2 \times 1.38 \times 10^{-23} (\text{JK}^{-1}) \times 300 (\text{K}) \times 0.693}{20 \times 1.67 \times 10^{-27} (\text{kg})} \right]^{\frac{1}{2}} \\ &\approx 8230 \text{ MHz}\end{aligned}$$

implying

$$\begin{aligned}\Delta\nu_D &= \frac{\Delta\omega_D}{2\pi} \\ &\approx 1300 \text{ MHz}\end{aligned}$$

where we have assumed

$$M_{\text{Ne}} \approx 20 M_H \approx 3.34 \times 10^{-26} \text{ kg}$$

The frequency variation of $g(\omega)$ is shown in Fig. 23.24. We may mention that for $\lambda_0 = 6328 \text{ \AA}$

$$\omega = \frac{2\pi c}{\lambda_0} \approx 2.98 \times 10^{15} \text{ s}^{-1}$$

Thus

$$\frac{\Delta\omega_D}{\omega} \approx 2.8 \times 10^{-6}$$

showing that the line-shape function is usually a very sharply peaked function. Further,

$$g(\omega_0) = \frac{2}{\Delta\omega_D} \left(\frac{\ln 2}{\pi} \right)^{\frac{1}{2}} \approx 1.1 \times 10^{-10} \text{ s} \quad (49)$$

(In Sec. 23.7.4 we will show that for a He-Ne laser, the Doppler broadening dominates over natural broadening and collisional broadening). If we assume a cavity with the following values of various parameters

$$d = 25 \text{ cm}, n_0 \approx 1, R_1 \approx 1, R_2 \approx 0.98, \alpha_c \approx 0$$

we would get

$$t_c \approx 8.2 \times 10^{-8} \text{ s}$$

Further, for the He-Ne laser

$$t_{sp} \approx 10^{-7} \text{ s}; n_0 \approx 1; \lambda_0 \approx 6328 \text{ \AA}$$

giving

$$(N_2 - N_1)_{th} \approx 3.7 \times 10^8 \text{ cm}^{-3}$$

For a given value of $(N_2 - N_1)$ (which is greater than the threshold value), a typical gain curve $\gamma(\nu)$ (which has a

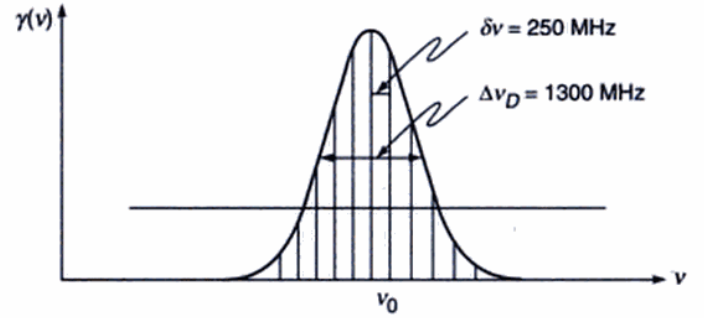


Fig. 23.25 For a given value of $(N_2 - N_1)$, a typical variation of the gain curve $\gamma(\nu)$. The vertical lines show the longitudinal modes of the cavity.

bandwidth of about 1300 MHz) is shown in Fig. 23.25. The horizontal line represents the value of

$$\frac{1}{(c/n_0)t_c} \quad (50)$$

For $n_0 \approx 1$, $t_c \approx 10^{-7} \text{ s}$ the above value is $\approx 3 \times 10^{-4} \text{ cm}^{-1}$. If we assume a 60 cm long He-Ne laser then the longitudinal mode spacing would be given by

$$\delta\nu = \frac{c}{2d} \approx 250 \text{ MHz} \quad (51)$$

and, as shown in Fig. 23.25, there will be seven longitudinal modes for which gain will exceed loss and which will oscillate. On the other hand, if d was only 10 cm then

$$\delta\nu \text{ would be } 1500 \text{ MHz}$$

and we will have single mode oscillation.

23.7 THE LINE-SHAPE FUNCTION

Since the line-shape function $g(\omega)$ determines the threshold population inversion [see Eq. (45)], we digress here to discuss some of the typical forms of $g(\omega)$ corresponding to different conditions.

We first consider the Doppler broadening which is due to the thermal motion of gas atoms. Also, in the He-Ne laser (which is probably the most popular laser), the line broadening mechanism is mainly due to Doppler broadening.

23.7.1 Doppler Broadening

In astronomy, we can determine how fast the stars or galaxies are moving (either directly away or directly towards us) by measuring the Doppler shift of spectral lines. For $v/c \ll 1$,

$$\omega - \omega_0 = \pm \omega_0 \frac{v}{c}; \quad (52)$$

the + sign corresponds to when the source of light is moving towards the observer and the - sign corresponds to when the source of light is moving away from the observer. Thus when the star is moving away from the observer, the measured frequency is slightly less than the actual value leading to the well-known *red shift* of spectral lines. Now, the probability that an atom has a z -component of velocity lying between v_z and $v_z + dv_z$ is given by the Maxwell distribution

$$P(v_z)dv_z = \left(\frac{M}{2\pi k_B T}\right)^{1/2} \exp\left(-\frac{Mv_z^2}{2k_B T}\right) dv_z \quad (53)$$

where M is the mass of the atom and T the absolute temperature of the gas. Notice that (using formula given in Appendix A):

$$\int_{-\infty}^{+\infty} P(v_z) dv_z = 1$$

as it indeed should be. Now, the probability $g(\omega) d\omega$ that the transition frequency lies between ω and $\omega + d\omega$ is equal to the probability that the z -component of the velocity of the atom lies between v_z and $v_z + dv_z$ where

$$v_z = \frac{(\omega - \omega_0)}{\omega_0} c \quad (54)$$

Thus

$$g(\omega)d\omega = \frac{c}{\omega_0} \left(\frac{M}{2\pi k_B T}\right)^{1/2} \exp\left[-\frac{Mc^2}{2k_B T} \frac{(\omega - \omega_0)^2}{\omega_0^2}\right] d\omega \quad (55)$$

which corresponds to a Gaussian distribution. The line-shape function is peaked at ω_0 , and the FWHM is given by

$$\Delta\omega_D = 2\omega_0 \left(\frac{2k_B T}{Mc^2} \ln 2\right)^{1/2} \quad (56)$$

where the subscript D implies that we are considering Doppler broadening. In terms of $\Delta\omega_D$ Eq.(55) can be written as

$$g(\omega)d\omega = \frac{2}{\Delta\omega_D} \left(\frac{\ln 2}{\pi}\right)^{1/2} \exp\left[-4 \ln 2 \frac{(\omega - \omega_0)^2}{(\Delta\omega_D)^2}\right] d\omega \quad (57)$$

A typical plot of the Gaussian line-shape function corresponding to the He-Ne laser is shown in Fig. 23.24.

23.7.2 Natural Broadening

The frequency spectrum associated with spontaneous emission is described by the Lorentzian line shape function

$$g(\omega) = \frac{1}{2\pi t_{sp}} \cdot \frac{1}{(\omega - \omega_0)^2 + \frac{1}{4t_{sp}^2}} \quad (58)$$

where

$$t_{sp} = \frac{1}{A_{21}} \quad (59)$$

represents the spontaneous emission lifetime. The FWHM of the Lorentzian is

$$\Delta\omega = \frac{1}{t_{sp}} = A_{21} \quad (60)$$

Thus, in terms of $\Delta\omega$, Eq. (58) can be written in the form

$$g(\omega) = \frac{\Delta\omega}{2\pi} \frac{1}{(\omega - \omega_0)^2 + \left(\frac{\Delta\omega}{2}\right)^2} \quad (61)$$

giving

$$g(\omega_0) = \frac{2}{\pi(\Delta\omega)} \quad (62)$$

Further

$$\int_0^{\infty} g(\omega) d\omega \approx \int_{-\infty}^{+\infty} g(\omega) d\omega = 1 \quad (63)$$

27.7.3 Collisional Broadening

In a gas, random collisions occur between the atoms. In such a collision process, when the atoms are very close to each other, the energy levels of the atoms change due to their mutual interaction. This leads to a Lorentzian line shape function given by

$$g(\omega) = \frac{\tau_0}{\pi} \frac{1}{1 + (\omega - \omega_0)^2 \tau_0^2} \quad (64)$$

where τ_0 represents the mean time between collisions;* The FWHM will be

$$\Delta\omega_c = \frac{2}{\tau_0}$$

In a typical gas laser $\tau_0 \sim 10^{-6}$ s giving

$$\Delta\omega_c \approx 2 \text{ MHz}$$

or

$$\Delta\nu_c \approx 0.3 \text{ MHz}$$

For the He-Ne laser, the Doppler line width is about 1300 MHz (see Example 23.3); on the other hand, the natural broadening is about 20 MHz and the collision broadening at 0.5 Torr is about 0.64 MHz. Thus, for He-Ne laser

*The derivation of Eq. (64) is given at many places, see, e.g. Sec. 8.8.2 of Ref. 20.

parameters, the Doppler broadening dominates over natural broadening and collision broadening.

The various line broadening mechanisms can be broadly classified under homogeneous and inhomogeneous broadening. Certain line broadening mechanisms, such as collision broadening or natural broadening, act to broaden the response of each atom in an identical fashion; such broadening mechanisms come under the class of homogeneous broadening. On the other hand, Doppler broadening or broadening produced due to local inhomogeneities in a crystal lattice act to shift the central frequency of the response of individual atoms by different amounts and thereby lead to an overall broadening of the response of the atomic system. Such a form of broadening is referred to as inhomogeneous broadening. If the effects which cause the inhomogeneous broadening are random in origin, then the broadened line is Gaussian in shape. In contrast, homogeneous broadening in general results in a Lorentzian line shape.

We return to Eq. (45) and notice that in order to have a low threshold value of population inversion:

- (i) The value of t_c should be large, i.e., the losses in the cavity must be small.
- (ii) The value of $g(\omega)$ at the centre of the line is $\approx 0.64/\Delta\omega$ for a Lorentzian line and $\approx 0.94/\Delta\omega$ for a Gaussian line (see Eqs 62 and 48). Thus, smaller the value of $\Delta\omega$ (the width of the line), smaller will be the threshold population inversion.
- (iii) Smaller values of t_{sp} (i.e., strongly allowed transitions) also lead to smaller values of threshold inversion. It must be noted here that, for shorter relaxation rates, larger pumping power is required to maintain a given amount of population inversion. In general, population inversion is more easily obtained on transitions which have longer relaxation times.
- (iv) The value of $g(\omega)$ at the centre of the line is inversely proportional to $\Delta\omega$, which, for example, in the case of Doppler broadening is proportional to ω [see Eq. (47)]. Thus, the threshold population inversion increases approximately in proportion to the third power of ω (apart from the frequency dependence of the other terms). Hence it is much easier to obtain laser action at infrared wavelengths than in the ultraviolet region.

23.8 TYPICAL PARAMETERS FOR A RUBY LASER

In order to get an idea of the magnitude of population inversion required for oscillation, we consider a ruby laser (see

Sec. 23.3). Let us consider the laser to be oscillating at the frequency corresponding to the peak of the emission line. We assume a concentration of 0.05 % of Cr^{3+} ions in the crystal; this corresponds to a population of

$$N = 1.6 \times 10^{19} \text{ Cr}^{3+} \text{ ions/cm}^3$$

For the case of ruby, the line is homogeneously broadened and the value of $g(\omega)$ at the peak of the line is $2 / (\pi\Delta\omega)$. Hence the threshold population inversion density is

$$\begin{aligned} (N_2 - N_1)_{th} &= \frac{\omega^2 n_0^3 t_{sp}}{\pi^2 c^3 t_c g(\omega)} \\ &= \frac{4\pi^2 n_0^3}{\lambda_0^3} \cdot \frac{\Delta\omega}{\omega} \cdot \frac{t_{sp}}{t_c} \end{aligned} \quad (65)$$

where λ_0 is the free space wavelength, t_{sp} is the spontaneous relaxation time of the upper laser level and t_c is the cavity lifetime. For ruby laser transition, one has

$$\begin{aligned} \lambda_0 &= 6943 \text{ \AA} \Rightarrow \omega = 2.715 \times 10^{15} \text{ s}^{-1} \\ \Delta\omega &\approx 9.4 \times 10^{11} \text{ s}^{-1}; \\ t_{sp} &= 3 \times 10^{-3} \text{ s}; n_0 = 1.76 \end{aligned}$$

where $n_0 (= 1.76)$ represents the refractive index of ruby. If we assume a cavity length of 5 cm and a loss per round trip of 10 % then $x = 0.1$ and using Eq. (40) we get

$$t_c = 6 \times 10^{-9} \text{ s}$$

Substituting all these values in Eq. (65) we get for the threshold population inversion density

$$(N_2 - N_1)_{th} \approx 1.1 \times 10^{17} \text{ Cr}^{3+} \text{ ions/cm}^3$$

Since the total density of Cr^{3+} ions in ruby is about $1.6 \times 10^{19} \text{ cm}^{-3}$, the fractional excess population required is very small.

We will next calculate approximately the minimum power required to maintain population inversion. Since t_{sp} represents the spontaneous relaxation time of the upper laser level, the number of atoms decaying per unit time from the upper laser level is approximately N_2/t_{sp} . For each atom lifted to level 2, one has to supply at least an amount of energy given by $h\nu_p$ where ν_p represents the average pump frequency. Hence in order to maintain N_2 atoms in the level 2, the minimum power P to be spent (per unit volume of the active material) would be given by

$$P = \frac{N_2 h\nu_p}{t_{sp}} \quad (66)$$

Now, since $(N_2 - N_1)_{th} \ll N$ (where N represents the total number of atoms per unit volume), we may write

$$N_2 \approx \frac{N}{2} \quad (67)$$

Thus, the minimum pumping power per unit volume required to maintain population inversion in a three level laser system is

$$P_{th} \approx \frac{N}{2} \frac{h\nu_p}{t_{sp}} \quad (68)$$

Taking the average pumping frequency as

$$\nu_p \approx 6.25 \times 10^{14} \text{ Hz}$$

(which is averaged over the green and violet absorption bands), we obtain

$$P_{th} \approx \frac{(1.6 \times 10^{19})}{2} \times \frac{6.6 \times 10^{-34} \times 6.25 \times 10^{14}}{3 \times 10^{-3}} \\ \approx 1100 \text{ W/cm}^3$$

If we assume that the efficiency of the pumping source is 25% and also that only 25% is absorbed in passage through the ruby rod, then the electrical threshold power comes out to be about 18 kW/cm³ of the active material. This is consistent with the threshold powers determined experimentally.

The threshold power calculation is particularly simple for the ruby laser where only three levels are involved. In general, in order to calculate the steady state population difference between the actual levels involved in the laser transition (for a given pumping rate) and also to know whether an inversion of population is achievable in a transition and if so, what would be the minimum pump power required to maintain a steady population inversion for continuous wave operation of the laser, it is necessary to solve equations which govern the rate at which populations of various levels change under the action of a pump and in the presence of laser radiation. These equations are referred to as 'rate equations' and have been discussed at many places; see, for example, References 4, 9, 16 and 17. We should mention that even for a three-level laser system the equation $N_2 = N/2$ [see Eq. (67)] is only approximately valid and in order to obtain a more accurate expression, it is necessary to solve the rate equations.

23.9 MONOCHROMATICITY OF THE LASER BEAM

Figure 23.26 shows the various line widths associated with a laser. The broad solid curve represents the spectral width

due to Doppler broadening of the laser medium. As an example if we consider the He-Ne laser operating at 6328 Å, the Doppler broadened linewidth is about 1,300 MHz. Inside the broad curve are shown the cavity modes as sharp peaks. The frequency separation between two adjacent cavity modes is $c/2d$ [see Eqs (6) and (51)] which for a typical laser cavity 60 cm long corresponds to 250 MHz; this is much less than the Doppler width (see Example 23.3). As we have discussed earlier, the cavity modes are also broadened due to the various losses in the cavity. Thus, for a 60 cm long cavity specified by a fractional loss per round trip of 4×10^{-2} , the width of the cavity mode is about 1.5 MHz. This is much smaller than the spacing between adjacent cavity modes. When the losses in the cavity are compensated for by the active medium placed inside the cavity, the resultant emission becomes extremely narrow and is limited due to the presence of spontaneous emission (which are random) and the fluctuations in the resonator parameters. The ultimate linewidth of an oscillating laser is determined solely by random spontaneous emissions can be shown to be given by*

$$(\delta\nu)_{sp} \approx \frac{2\pi(\Delta\nu_p)^2 h\nu_0}{P^0}$$

where ν_0 is the frequency of oscillation, P^0 is the output power and

$$\Delta\nu_p = \frac{1}{2\pi t_c}$$

is known as the passive cavity linewidth, t_c being the cavity lifetime (see Sec. 23.6.2). The subscript 'sp' refers to the fact that the linewidth is due to spontaneous emissions. The decrease in $(\delta\nu)_{sp}$ with increase in power output is due to the fact that for a given mirror transmittance, increase in P^0

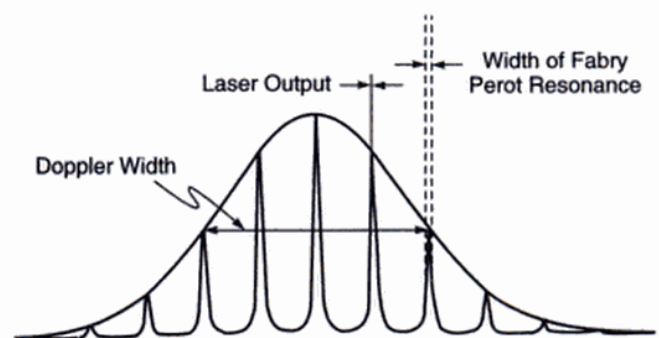


Fig. 23.26 The solid curve represents a typical Doppler broadened spectral line. The closely spaced cavity modes are shown as narrow peaks inside the curve. The sharp line represents the output of the laser (Ref. 21).

*See., e.g., Ref. 22.

corresponds to increase in laser power inside the resonator cavity, and this leads to the dominance of stimulated emissions over spontaneous emission

As a typical example, $\Delta\nu_p \approx 1 \text{ MHz}$, $P^0 = 1 \text{ mW} = 10^{-3} \text{ W}$, $h\nu = 2 \times 10^{-19} \text{ J}$ (corresponding to the red region of the spectrum) so that $(\delta\nu)_{sp} \approx 10^{-3} \text{ Hz}$ —an extremely small quantity indeed! Thus the ultimate monochromaticity is determined by the spontaneous emissions occurring inside the cavity because the radiation coming out due to spontaneous emission is incoherent. However, in practice, the monochromaticity is limited by external factors like temperature fluctuations, mechanical vibrations of the optical cavity, etc. For example, if we assume the oscillation frequency of a mode is given by Eq. (4) then, the change in frequency $\Delta\nu$ caused by a change in length Δd is given by

$$\frac{\Delta\nu}{\nu} = \frac{\Delta d}{d}$$

Thus for $d \approx 50 \text{ cm}$, if we assume a stability of $\Delta d \approx 1 \text{ \AA}$ then for $\nu \approx 5 \times 10^{14} \text{ Hz}$:

$$\Delta\nu \approx 10^5 \text{ Hz}$$

which is much much larger than $(\delta\nu)_{sp}$. We may mention here that $\Delta\nu \approx 10^5 \text{ Hz}$ correspond to $\Delta\lambda \approx 10^{-6} \text{ \AA}$. Indeed, for a single moded He-Ne laser, we can have $\Delta\nu \approx 10^5 \text{ Hz}$. On the other hand, for a multimoded He-Ne laser $\Delta\lambda \sim 0.02 \text{ \AA}$ implying a coherence length of about 20 cm .

23.10 RAMAN AMPLIFICATION AND RAMAN LASER

We will first discuss the physics of Raman effect. When a monochromatic light beam gets scattered by a transparent substance, one of the following may occur:

1. Over 99% of the scattered radiation has the same frequency as that of the incident light beam; this is known as Rayleigh scattering which has been discussed in Sec. 6.6. The sky looks blue because of Rayleigh scattering and the light that comes out from the side of the optical fiber (see Fig. 24.2) is also due to Rayleigh scattering.
2. A very small portion of the scattered radiation has a frequency different from that of the incident beam—this may arise due to one of the following three processes:

- (i) The incident radiation may lead to translatory motion of the molecules—this would result in

shift of frequency which is usually very small and difficult to measure. This is known as Brillouin scattering*.

- (ii) A part of the energy $h\nu$ of the incident photon is taken over by the molecule in the form of rotational (or vibrational) energy and the scattered photon has a smaller energy $h\nu'$. This leads to what are known as Raman – Stokes lines [see Fig. 23.10(a) and 23.11].
- (iii) On the other hand, the photon can undergo scattering by a molecule which is already in an excited state. The molecule can de-excite to one of the lower energy states and in the process, the incident photon can take up this excess energy and come out with a higher frequency. This leads to what are known as Raman anti-Stokes lines [see Fig. 23.10(b) and 23.11].

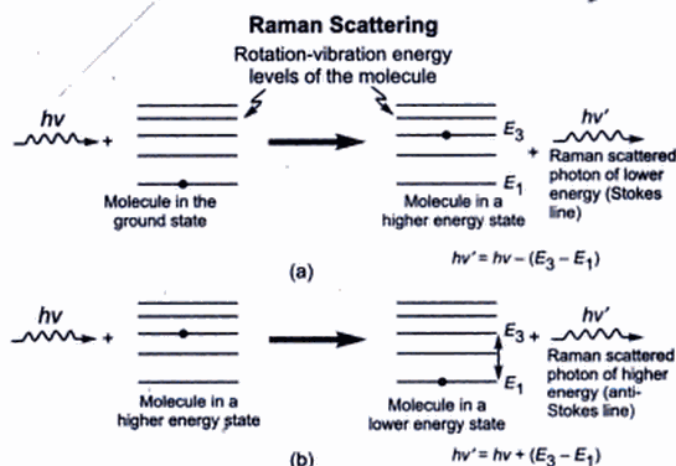


Fig. 23.10 The generation of the Raman Stokes and the Raman anti-Stokes lines.

The difference energy, which is $(h\nu - h\nu')$ for the Raman Stokes line and $(h\nu' - h\nu)$ for the Raman anti-Stokes line, would therefore correspond to the energy difference between the rotational (or vibrational) energy levels of the molecule and would therefore be a characteristic of the molecule itself.

The quantity $(h\nu - h\nu')$ or $(h\nu' - h\nu)$ is usually referred to as the "Raman-shift" [see Fig. 23.11] and is independent of the frequency of the incident radiation. Through a careful analysis of the Raman spectra, one can determine the structure of molecules; there lies the tremendous importance of Raman effect. The intensity distribution of a typical Raman spectrum for the CCl_4 molecule is shown in Fig. 23.11.

*The shift in frequency is usually denoted by $\Delta\bar{\nu}$ and is measured in wavenumber units which is defined later in this section. In Brillouin scattering, $\Delta\bar{\nu} \leq 0.1 \text{ cm}^{-1}$. On the other hand, in Raman scattering $\Delta\bar{\nu} \leq 10^4 \text{ cm}^{-1}$.

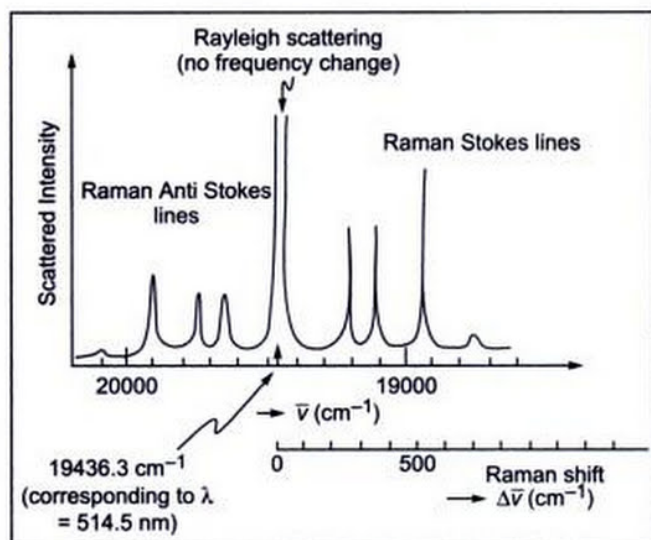


Fig. 23.11 Raman spectra of CCl_4 excited by 514.5 nm line of an Argon-ion laser [Adapted from <http://epsc.wustl.edu/Haskin-group/Raman/faqs.htm>]

In spectroscopy, the energy levels of atoms or molecules and also the energy of a photon are measured in wavenumber units which are obtained by dividing the energy by hc , where h ($\approx 6.56 \times 10^{-27}$ ergs-s) is the Planck's constant and c ($\approx 3 \times 10^{10}$ cm/s) is the speed of light in free space — in spectroscopy everyone uses CGS units!! In the case of molecular (or atomic) energy levels, these are usually denoted by the symbol T_n :

$$T_n = \frac{E_n}{hc}$$

The photon's energy is $h\nu$ and therefore, in wavenumber units

$$\frac{h\nu}{hc} = \frac{\nu}{c} = \frac{1}{\lambda}$$

is just the inverse of the wavelength and is usually denoted by the symbol $\bar{\nu}$. Thus

$$\bar{\nu} = \frac{1}{\lambda}$$

Now, the energy levels of the hydrogen atom in wavenumber units are given by

$$T_n = \frac{E_n}{hc} = -\frac{R}{n^2}; n = 1, 2, 3, \dots$$

where R ($\approx 109678 \text{ cm}^{-1}$) is known as the Rydberg constant and n ($=1, 2, 3, \dots$) is the total quantum number of the state. Thus corresponding to the $n = 3$ to $n = 2$ transition (one of the lines of the Balmer series) we will get a photon of wavenumber

$$\bar{\nu} = -R\left(\frac{1}{9} - \frac{1}{4}\right) = \frac{5}{36} \times 109678 = 15233 \text{ cm}^{-1}$$

The inverse of the above number ($\approx 6.56 \times 10^{-5} \text{ cm}$) represents the wavelength of the emitted photon.

Figure 23.11 shows the intensity distribution of the Raman spectrum of CCl_4 molecule when the incident radiation corresponds to the Argon-ion laser line having a wavelength of $5.145 \times 10^{-5} \text{ cm}$; in wavenumber units the value is 19436.3 cm^{-1} . The central peak in the figure corresponds to this wavelength and is due to Rayleigh scattering. The Raman shift for the Stokes lines is the same as for the anti-Stokes lines although the latter is much weaker. This is due to the fact that at room temperature, the number of molecules in the ground state is much larger than the molecules present in excited states. This leads to very low intensities of the Raman anti-Stokes lines. The actual Raman spectrum of the CCl_4 molecule for the 4046 Å lines of mercury lamp is shown in Fig. 23.12. The photograph is adapted from the 1930 Nobel lecture of C.V. Raman. It may be of interest to mention that on 28th February 1928, K.S. Krishnan and C.V. Raman observed "Raman effect" in several organic vapours like pentane — which they called "the new scattered radiation". Raman made a newspaper announcement on 29th February and on 8th March 1928, he communicated a paper entitled "A Change of Wavelength in Light Scattering" to *Nature*; the paper was published on 21st April 1928. Although in the paper, he acknowledged that the observations were made by K.S. Krishnan and himself, the paper had Raman as the author and therefore the phenomenon came to be known as Raman effect although many scientists (particularly in India) kept on referring it as the Raman-Krishnan effect. Subsequently, there were several papers written by Raman and Krishnan. Raman got the Nobel prize in 1930 for "his work on the scattering of light and for the discovery of the effect named after him". At about the same time, Landsberg and Mandel'shtam (in Russia) were also working on light scattering and according to Mandel'shtam, they observed the "Raman-lines" on February 21, 1928. But the results were presented in April 1928 and it was only on 6th May 1928



Fig. 23.12 The observed Raman spectra of CCl_4 for the 4046 Å and 4358 Å lines of mercury lamp. The photograph is adapted from the 1930 Nobel lecture of C.V. Raman.

that Landsberg and Mandel'shtam communicated their results to the journal *Naturwissenschaften*. But by then it was too late! Much later, scientists from Russia kept calling Raman scattering as Mandel'shtam-Raman scattering. For a very nice historical account of Raman effect, we refer the reader to a book by G. Venkataraman on "Journey into Light: Life and Science of C.V. Raman" published by Penguin Books (1994).

In 1958, thirty years after the discovery of the Raman effect, Raman wrote an article on "Raman effect" in Encyclopaedia Britannica. In that article he wrote "The rotations of the molecules in gases give more readily observable effects, viz., a set of closely spaced but nevertheless discrete Raman lines located on either side of the incident line. In liquids, only a continuous wing or band is usually observed in the same region, indicating that the rotations in a dense fluid are hindered by molecular collisions. The internal vibrations of the molecules, on the other hand, give rise in all cases to large shifts of wavelength. The Raman lines attributed to them appear well separated from the parent line and are therefore easily identified and measured."

In stimulated Raman emission, the radiation emitted in the ordinary Raman effect is made to stimulate further Raman emission. This can lead to what is usually referred to as the "Raman amplification" of the beam.

Now, in fused silica, because of interaction between adjacent SiO_2 molecules, the vibrational bands are very broad; this leads to a very broad Raman shift lying between 430 cm^{-1} to 470 cm^{-1} [this corresponds to a Raman frequency shift between 13 and 14 THz ($1\text{ THz} = 10^{12}\text{ Hz}$)]. Thus if we have a pump laser at 1450 nm ($\bar{\nu} = 6897\text{ cm}^{-1}$)

then an incoming beam at 1550 nm ($\bar{\nu} = 6452\text{ cm}^{-1}$) will get amplified by stimulated Raman scattering [$\Delta\bar{\nu} = 445\text{ cm}^{-1}$] as shown in Fig. 23.13 (a). In an actual commercially available single mode fiber of length about 30 km, one can obtain a Raman gain of about 15 dB (i.e., a power amplification by a factor of about 30) by using a pump laser of 500 mW power.

Similarly, if we want to amplify an incoming beam at 1300 nm ($\bar{\nu} = 7692\text{ cm}^{-1}$) then we must use a pump laser at about 1230 nm wavelength ($\bar{\nu} = 8130\text{ cm}^{-1}$) as shown in Fig. 23.13 (b). This is the great advantage of the Raman fiber amplifier. One can amplify signal at *any* wavelength provided we choose the pump laser frequency separated by about 13.5 THz (equivalent to a wavenumber shift of about 450 cm^{-1}). On the other hand, as we may recall, in Erbium Doped Fiber Amplifiers (EDFAs), one can amplify signals only around 1550 nm wavelength; however, the laser power required is much smaller.

The above principle can be used to build the cascaded Raman laser (see Fig. 23.14). The vertical bars represent FBGs (Fiber Bragg Gratings) which are strongly reflecting at the wavelengths written on the top (see Sec. 13.4.1 for a brief account on FBGs). Thus the input wavelength of 1100 nm ($\approx 9091\text{ cm}^{-1}$) produces Raman scattered line at 1155 nm ($\approx 8658\text{ cm}^{-1}$) implying a Raman shift of about 433 cm^{-1} ; this resonates between two FBGs having peak reflectivity at 1155 nm . Now, this 1155 nm ($\approx 8658\text{ cm}^{-1}$) beam produces Raman scattered line at 1218 nm ($\approx 8210\text{ cm}^{-1}$) implying a Raman shift of about 448 cm^{-1} which resonates between two FBGs having peak reflectivity at 1218 nm etc. This way laser output can be generated anywhere from $1100 - 1600\text{ nm}$ [see Fig. 23.14].

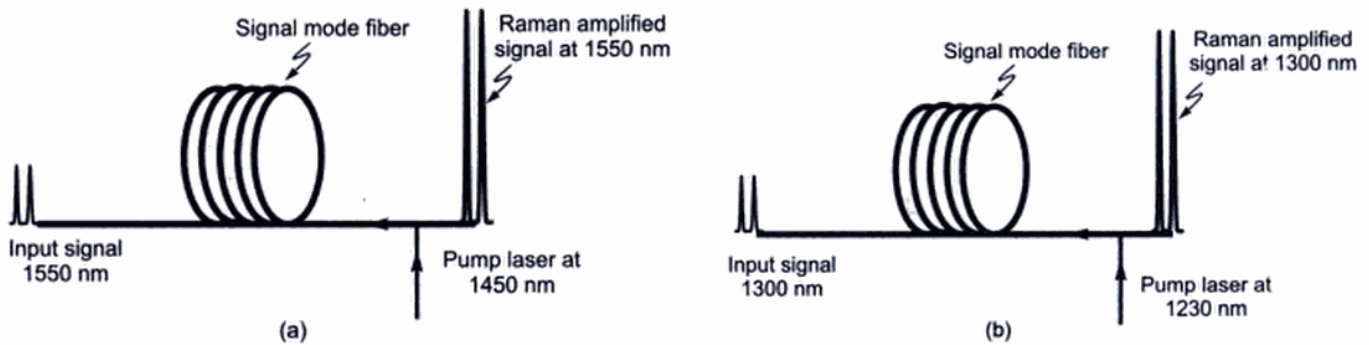


Fig. 23.13 Raman fiber amplifiers at 1550 nm and 1300 nm wavelengths.

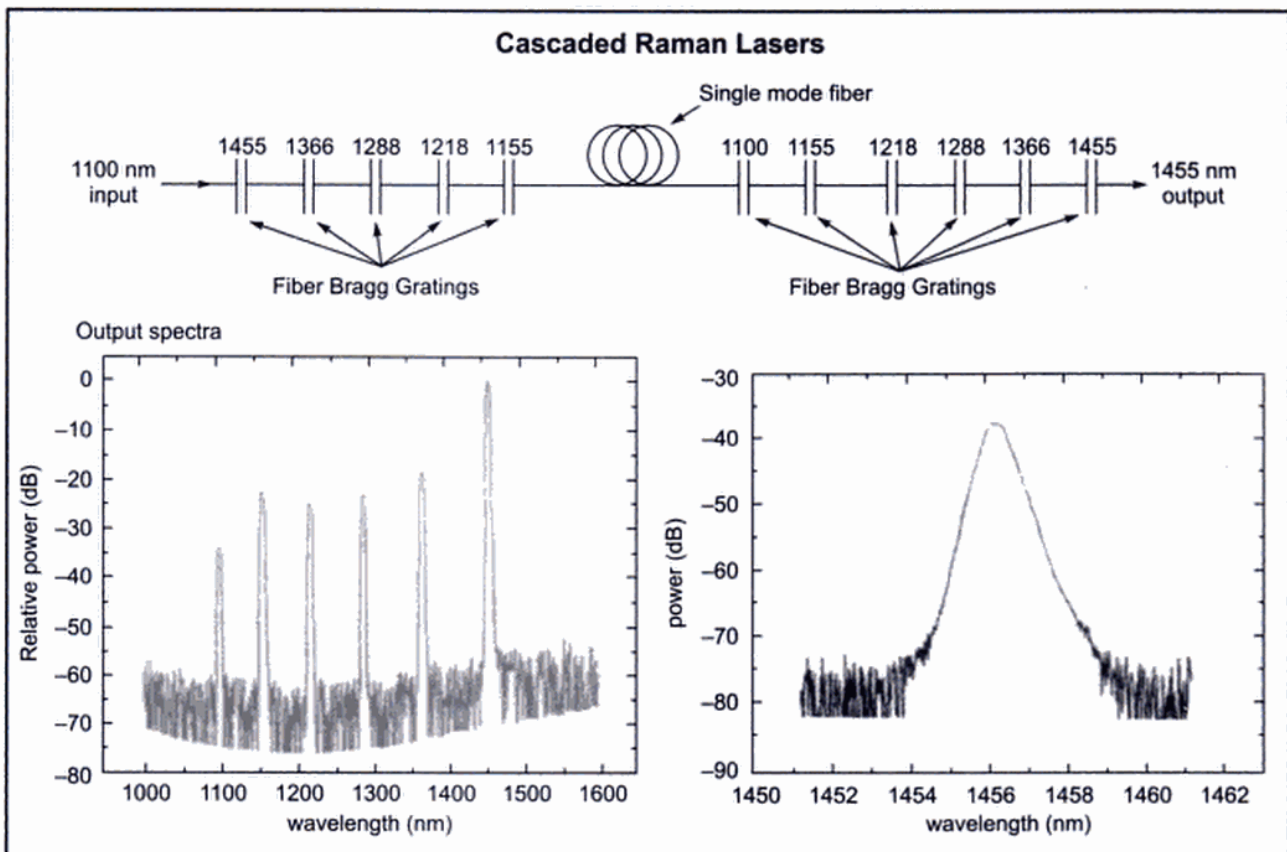


Fig. 23.14 The cascaded Raman laser; output can be generated anywhere from 1100 - 1600 nm. [Adapted from the lecture notes of K. Rottwitt on "Raman Amplification using Optical Fibers", CGCRI, Kolkata].

SUMMARY

- Laser is an acronym for Light Amplification by Stimulated Emission of Radiation. The light emitted from a laser often possesses some very special characteristics - some of these are (a) **Directionality**: because of which a laser beam can be focused to areas \sim few $(\mu\text{m})^2$ leading to applications in surgery, material processing, compact discs, etc. (b) **High power**: continuous wave lasers having power levels $\sim 10^5$ W and pulsed lasers having a total energy ~ 50000 J have applications in welding, cutting, laser fusion etc. and (c) **Spectral purity**: Laser beams can have an extremely small spectral width $\Delta\lambda$, because of which lasers find applications in holography, optical communications, spectroscopy etc.
- As put forward by Einstein, when an atom is in the excited state then, in addition to the spontaneous emission, it can also make a transition to a lower energy state by what is known as stimulated emission in which an incident signal of appropriate frequency triggers an atom in an excited state to emit radiation

- this results in the amplification of the incident beam. If we are able to create a state of population inversion in which there are larger number of atoms in the upper state then the number of stimulated emissions would exceed the number of stimulated absorptions resulting in the (optical) amplification of the beam.

- The three main components of any laser are
 - (i) The active medium which consists of a collection of atoms, molecules or ions (in solid, liquid or gaseous form), which is capable of amplifying light waves,
 - (ii) The pumping mechanism which allows us to obtain a state of population inversion between a pair of energy levels of the atomic system, and
 - (iii) The optical resonator, which provides the feed back.
- Through a pumping mechanism, one creates a state of population inversion in the laser placed inside the resonator system. The spontaneous emission occurring inside the resonator cavity excites the various modes of the cavity. The modes for which the gain is

higher than the losses get amplified by drawing energy from the laser medium. The amplitude of the mode increases rapidly until the upper level population reaches a value when the gain equals the losses and the mode oscillates in steady state.

- Two mirrors facing each other form a resonant cavity. The discrete frequencies of the resonator modes are given by $\nu = \nu_m = m \frac{c}{2d}$. Different values of m lead to different oscillation frequencies, which constitute the longitudinal modes of the cavity. For example, for an optical resonator of length $d \approx 60$ cm operating at an optical frequency of $\nu \approx 5 \times 10^{14}$ Hz (corresponding to $\lambda \approx 6000$ Å), we obtain $m \approx 2 \times 10^6$.
- The first successful operation of a laser device ($\lambda \sim 0.684$ μm) was demonstrated by Theodore Mainman in 1960 using a ruby crystal. Within a few months of the operation of the ruby laser, Ali Javan and his associates constructed the first gas laser ($\lambda \sim 0.633$ μm), namely the helium – neon laser.
- If we put a fiber (doped with Erbium or Neodymium) between two mirrors (which act as a resonator)—then with an appropriate pump we would have a fiber laser. In 1961, the first fiber laser (barium crown glass doped with Nd^{+3} ions) was fabricated by Elias Snitzer.
- The threshold population inversion required for the oscillation of the laser is given by

$$(N_2 - N_1)_{\text{th}} = \frac{\omega^2 n_0^3 t_{\text{sp}}}{\pi^2 c^3 t_c g(\omega)}$$

where t_{sp} is the spontaneous emission life time, t_c is the passive cavity life time and $g(\omega)$ is the line shape function. For a He – Ne laser

$$g(\omega) = \frac{2}{\Delta\omega_D} \left(\frac{\ln 2}{\pi} \right)^{\frac{1}{2}} \exp \left[-4 \ln 2 \frac{(\omega - \omega_0)^2}{\Delta\omega_D^2} \right]$$

where $\Delta\omega_D = 2\omega_0 \left(\frac{2k_B T}{Mc^2} \ln 2 \right)^{\frac{1}{2}}$ represent the FWHM

(Full Width at Half Maximum) of the line k_B the Boltzmann constant, T represents the absolute temperature of the gas and M represents the mass of the atom responsible for the lasing transition (neon in the case of a He – Ne laser). Notice that the minimum threshold value of $N_2 - N_1$ would correspond to the center of the line where $g(\omega)$ is a maximum and for He – Ne laser at $T = 300^\circ$ K, $\Delta\omega_D \approx 8230$ MHz giving $g(\omega_0) \approx 1.1 \times 10^{-10}$ s. Assuming $M = 20 M_H \approx 3.3 \times 10^{-23}$ g, $t_c \approx 10^{-7}$ s $\approx t_{\text{sp}}$, $n_0 \approx 1$, we get $(N_2 - N_1)_{\text{th}} \approx 4 \times 10^8 \text{ cm}^{-3}$.

PROBLEMS

- 23.1 Determine the MKS units of $u(\omega)$, u_ω , A and B .

[Ans. J s m^{-3} ; J m^{-3} ; s^{-1} ; $\text{m}^3 \text{J}^{-1} \text{s}^{-2}$].

- 23.2 For the $2P \rightarrow 1S$ transition in the hydrogen atom calculate ω . Assuming the spontaneous emission lifetime of the $2P$ state to be 1.6 ns, calculate the Einstein B coefficient. Assume $n_0 \approx 1$.

[Ans. $\omega \approx 1.5 \times 10^{16}$ Hz,
 $B_{21} \approx 4.2 \times 10^{20} \text{ m}^3 \text{J}^{-1} \text{s}^{-2}$]

- 23.3 (a) Consider a He–Ne laser with cavity life time $t_c \approx 5 \times 10^{-8}$ s. If $R_1 = 1.0$ and $R_2 = 0.98$, calculate the cavity length d ; assume $n_0 \approx 1$.
(b) Calculate $\Delta\nu_p$ and compare with the longitudinal mode spacing $\delta\nu$.

[Ans. (a) $d \approx 15$ cm
(b) $\Delta\nu_p \approx 3.2$ MHz; $\delta\nu \approx 1$ GHz]

- 23.4 In a typical He–Ne laser ($\lambda = 6328$ Å) we have $d \approx 20$ cm, $R_1 = R_2 = 0.98$, $\alpha_c \approx 0$, $t_{\text{sp}} \approx 10^{-7}$ s, $\Delta\nu_D \approx 1.3 \times 10^9$ Hz and $n_0 \approx 1$. Calculate t_c and $(N_2 - N_1)_{\text{th}}$.

[Ans. 33 ns; $8.8 \times 10^8 \text{ cm}^{-3}$]

- 23.5 Consider the D_1 line of Na ($\lambda \approx 5890$ Å)

(a) The spontaneous emission lifetime $t_{\text{sp}} \approx 16$ ns. Calculate $\Delta\nu_N$ and $\Delta\lambda_N$.

(b) Assume $T = 500^\circ$ K. Calculate $\Delta\nu_D$ and $\Delta\lambda_D$.
[$k_B \approx 1.38 \times 10^{-23} \text{ J/}^\circ\text{K}$; $M_{\text{Na}} \approx 23 M_H$; $M_H \approx 1.67 \times 10^{-27} \text{ kg}$].

[Ans. $\Delta\lambda_N \approx 10^{-4}$ Å; $\Delta\lambda_D \approx 0.02$ Å]

- 23.6 In a CO_2 laser ($\lambda_0 = 10.6$ μm), the laser transition occurs between the vibrational states of the CO_2 molecule. At $T \approx 300^\circ$ K, calculate the Doppler linewidth $\Delta\nu_D$ and $\Delta\lambda_D$ [$M_{\text{CO}_2} \approx 44 M_H$].

[Ans. $\Delta\nu_D \approx 53$ MHz; $\Delta\lambda_D \approx 0.2$ Å]

- 23.7 Consider a light beam of all frequencies lying between $\nu = \nu_0 = 5.0 \times 10^{14}$ Hz to $\nu = 5.00002 \times 10^{14}$ Hz incident normally on a resonator (see Fig. 23.16) with $R = 0.95$, $n_0 = 1$ and $d = 25$ cm. Calculate the frequencies (in the above frequency range) and the corresponding mode number which will correspond to transmission resonances.

[Ans. $\nu = \nu_0 + 400$ MHz ($m = 833,334$),
 $\nu_0 + 1000$ MHz ($m = 833,335$)
and $\nu_0 + 1600$ MHz ($m = 833,336$)]

- 23.8 Referring to Fig. 23.9, if $d = 2R_1 = 2R_2$ show that all rays passing through the common center of curvature of the mirrors will retrace their path and hence be trapped inside the cavity.

- 23.9** Consider a He-Ne laser ($\lambda_0 = 0.6328 \mu\text{m}$) with $d = 30 \text{ cm}$, $n_0 \approx 1$, $R_1 \approx 1$, $R_2 \approx 0.99$. Calculate the passive cavity linewidth $\Delta\nu_p$ and the passive cavity life time t_c . You may assume $\alpha_c \approx 0$.

[Ans. 0.8 MHz, 0.2 μs]

- 23.10** (a) For the He-Ne laser described in Problem 23.9, if the power level is 0.5 mW, calculate the ultimate linewidth $(\delta\nu)_{sp}$.
(b) Discuss the stability of the mirror position Δd to obtain the ultimate linewidth.

REFERENCES AND SUGGESTED READINGS

1. C. H. Townes, 'Production of Coherent Radiation by Atoms and Molecules' in *Nobel Lectures in Physics (1963–1970)*, Elsevier Publishing Company, Amsterdam, 1972. (Reprinted in Ref. 4).
2. N. G. Basov, 'Semiconductor Lasers' in *Nobel Lectures in Physics (1963–1970)*, Elsevier Publishing Company, Amsterdam, 1972. (Reprinted in Ref. 4).
3. A. M. Prochorov, 'Quantum Electronics' in *Nobel Lectures in Physics (1963–1970)*, Elsevier Publishing Company, Amsterdam, 1972. (Reprinted in Ref. 4).
4. K. Thyagarajan and A. K. Ghatak, *Lasers: Theory and Applications*, Plenum Press, New York, 1981. (Reprinted by Macmillan India, New Delhi).
5. A. Einstein, 'On the Quantum Theory of Radiation', *Physikalische Zeitschrift*, Vol. 18, 121, 1917 [Reprinted in Ref. 6].
6. D. Ter Haar, *The Old Quantum Theory*, Pergamon Press, Oxford, 1967.
7. A. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, 1998.
8. E. Desurvire, *Erbium Doped Fiber Amplifiers*, John Wiley, New York, 1994.
9. A. E. Siegman, *Lasers*, Oxford University Press, Oxford, 1986.
10. W. Johnstone, *Erbium Doped Fiber Amplifiers*, Unpublished lecture notes.
11. S. Yoshida, S. Kuwano and K. Iwashita, 'Gain Flattened EDFA with high Al-concentration for multistage repeated WDM transmission experiments', *Electronics Letters*, Vol. 31, 1765, 1995.
12. E. Snitzer, 'Optical Maser Action of Nd^{+3} in a Barium Crown Glass', *Physical Review Letters*, Vol. 7, 444–446, 1961.
13. T. H. Maiman, 'Stimulated Optical Radiation in Ruby', *Nature*, Vol. 187, 493–494, 1960.
14. R. Brown, *Lasers, A Survey of Their Performance and Applications*, Business Books, London, 1969.
15. A. Javan, W. R. Bennett Jr. and D. R. Herriott, 'Population Inversion and Continuous Optical Maser Oscillation in a Gas Discharge Containing a He-Ne Mixture', *Physical Review Letters*, Vol. 6, 106–110, 1961.
16. C. C. Davis, *Lasers and Electro-Optics*, Cambridge University Press, Cambridge, 1996.
17. J. T. Verdeyen, *Laser Electronics*, Prentice-Hall, Englewoodcliffs, N. J., 1989.
18. C. Lin, 'Optical Communications: single mode optical fiber transmission systems', in *Optoelectronic Technology and Lightwave Communication Systems*, Ed. C. Lin, Van Nostrand Reinhold, New York, 1989.
19. A. K. Ghatak and S. Lokanathan, *Quantum Mechanics*, 5th Edition, Macmillan India, New Delhi, 2004. [Also published by Kluwer Academic Publishers, Dordrecht, 2004].
20. A. K. Ghatak and K. Thyagarajan, *Optical Electronics*, Cambridge University Press, Cambridge, 1989.
21. D. R. Herriot, 'Optical Properties of a Continuous He-Ne Optical Maser', *Journal of the Optical Society of America, USA*, Vol. 52, p. 31, 1962.
22. A. Maitland and M. H. Dunn, *Laser Physics*, North Holland Publishing Co., Amsterdam, 1969.

PART 8

Some Contemporary Topics

This part consists of two chapters — the first chapter is an introduction to the basics of fiber optics discussing specially the characteristics of optical fibers as regards to their application to telecommunications and fiber optic sensors. Following a brief historical introduction, the basic principle of light guidance in an optical fiber is discussed followed by a detailed analysis of two of its important characteristics: attenuation and pulse dispersion. The second chapter is a very short chapter — essentially an essay on Speckle metrology.

room temperature operation of semiconductor lasers in 1970. Ever since, the scientific and technological progress in this field has been so phenomenal that we are already in the 5th generation of optical fiber communication systems within a brief span of 30 years. Recent developments in optical amplifiers and WDM (wavelength division multiplexing) are taking us to a communication system with almost 'zero' loss and almost 'infinite' bandwidth. Indeed optical fiber communication systems are fulfilling the increased demand on communication links especially with the proliferation of the internet.

This chapter is an introduction to the basics of fiber-optics discussing specially the characteristics of optical fibers as regards to their application to telecommunications and fiber-optic sensors. Following a historical introduction, we will discuss the basic principle of light guidance in an optical fiber and also its two important characteristics: attenuation and pulse dispersion. We will also briefly discuss simple fiber-optic sensors and also plastic optical fibers. Fiber amplifiers and fiber lasers have been very briefly discussed in the previous chapter.

24.2 SOME HISTORICAL REMARKS

Communication implies transfer of information from one point to another. When it is required to transmit some information such as speech, images, data, etc. over a distance, one generally uses the concept of carrier wave communication. In such a system, the information to be sent modulates an electromagnetic wave such as radio wave, or microwave which acts as a carrier. This modulated wave is then transmitted to the receiver through a channel and the receiver receives the modulated wave and demodulates it to retrieve the signal. For example, the AM broadcast usually range from about 600 kHz to about 2 MHz; the abbreviation AM wave stands for an amplitude modulated wave. If we assume that the highest frequency associated with music is about 20 kHz ($= 0.02$ MHz), then at a carrier frequency of 1.5 MHz, the frequency of the AM wave must vary between 1.48 MHz to 1.52 MHz—a bandwidth of 40 kHz. Thus in the entire AM broadcast range from about 600 kHz to about 2 MHz we can have at most about 30 channels; indeed we will have less number of channels if we use more bandwidth for each channel. On the other hand in TV transmission since we have to scan pictures, more information needs to be sent and we require much greater bandwidth (about 5 MHz) — necessitating higher carrier frequency; the carrier

frequencies associated with the TV broadcast range from about 500 MHz to about 900 MHz.

Since optical beams have frequencies in the range of 10^{14} – 10^{15} Hz, the use of such beams as the carrier would imply a tremendously large increase in the information transmission capacity of the system as compared to systems employing radio waves or micro waves. It is this large information carrying capacity of a light beam that has generated interest amongst communication engineers to develop a communication system using light waves as carrier waves.

Now, in a conventional telephone hook up, voice signals are converted into equivalent electrical signals by the microphone and are transmitted as electrical currents through metallic (copper or aluminum) wires to the local telephone exchange. Thereafter, these signals continue to travel as electric currents through metallic wire cable (or for long distance transmission as radio/microwaves to another telephone exchange) usually with several repeaters in between. From the local area telephone exchange at the receiving end these signals travel to the receiver telephone via metallic wire pairs where they are converted back into corresponding sound waves. Through such cabled wire-pair telecommunication systems, one can at most send 48 simultaneous telephone conversations intelligibly. On the other hand in an optical communication system, which utilizes glass fibers as the transmission medium and lightwaves as carrier waves, it has been possible (in 2001) to send over 1 terabit of information in a second (which is roughly equivalent to transmission of about 15 million simultaneous telephone conversations) through one hair thin optical fiber. This is certainly one of the extremely important technological achievements of the twentieth century.

The idea of using light waves for communication can be traced to as far back as 1880 when Alexander Graham Bell invented the photophone (see Fig. 24. 1) shortly after he invented the telephone* in 1876. In this remarkable experiment, speech was transmitted by modulating a light beam, which travelled through air to the receiver. The transmitter consisted of a flexible reflecting diaphragm which could be activated by sound and which was illuminated by sunlight. The reflected light was collimated by a lens and the reflected beam was received by a parabolic reflector placed at a distance. The parabolic reflector concentrated the light on a photoconducting selenium cell, which forms a part of a circuit with a battery and a receiving earphone. Sound waves present in the vicinity of the diaphragm vibrate the diaphragm which leads to a consequent variation of the light

*Actually according to newspaper reports (published in June 2002), an Italian immigrant Antonio Meucci, was the inventor of telephone. According to this report, Antonio Meucci demonstrated his 'teletrfeno' in New York in 1860. Alexander Graham Bell took out his patent 16 years later. This has apparently been recognized by US Congress.

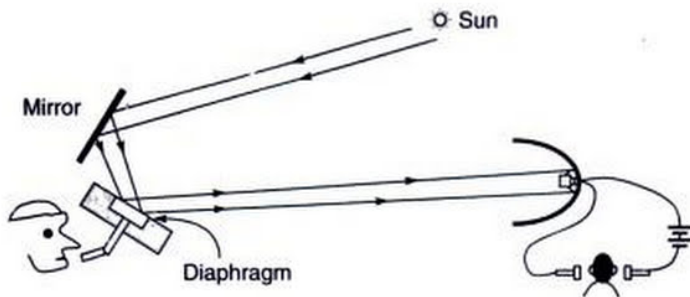


Fig. 24.1 Schematic of the photophone invented by Graham Bell. In this system, sunlight was modulated by a diaphragm and transmitted through a distance of about 200 meters in air to a receiver containing a selenium cell connected to the earphone.

reflected by the diaphragm. The variation of the light falling on the selenium cell changes the electrical conductivity of the cell, which in turn changes the current in the electrical circuit. This changing current reproduces the sound on the earphone. To quote from Ref. 3:

In 1880 he (Graham Bell) produced his 'photophone' which to the end of his life, he insisted was '... the greatest invention I have ever made, greater than the telephone ...'. Unlike the telephone it had no commercial value.

After the photophone experiment of Graham Bell there was almost no serious effort in developing optical communication systems. This was primarily due to the fact that earlier, there were no suitable light sources available that could reliably be used as the information carrier. Then a revolution occurred—the first laser was fabricated in 1960. Although incoherent sources like LED's (light emitting diodes) are also sometimes used in present day optical communication systems, it was the discovery of the laser which triggered serious interest, for the first time, in the development of optical communication systems.

On the other hand, around the same time telecommunications traffic was growing so rapidly that it was felt that conventional telecommunication systems based on, say, coaxial cables, radio and microwave links, and wire-pair cable, could soon reach a saturation point. The advent of lasers thus immediately triggered a great deal of investigations aimed at examining the possibility of building optical analogues of conventional communication systems. The very first such modern optical communication experiments involved laser beam transmission through the atmosphere. However, it was soon realized that laser beams could not be

sent in open atmosphere through reasonably long distances to carry signals unlike, for example, microwave or radio systems operating at longer wavelengths. This is due to the fact that a light beam (of wavelength about $1\ \mu\text{m}$) is severely attenuated and distorted owing to scattering and absorption by the atmosphere. Thus for reliable lightwave communication under terrestrial environments it would be necessary to provide a guiding medium that can protect the signal carrying light beam from the vagaries of the terrestrial atmosphere. This guiding medium is the optical fiber which is a hair-thin structure and guides the light beam from one place to another (see Fig. 24.2); the guidance of the light beam through the optical fiber takes place because of the phenomenon of total internal reflection which we will discuss in the following section.



Fig. 24.2 A long and thin optical fiber transmitting a light beam. (After Ref. 4; photograph courtesy Dr. Chynoweth).

In addition to the capability of carrying a huge amount of information, optical fibers fabricated with recently developed technology are characterized by extremely low losses** ($< 0.25\ \text{dB/km}$) as a consequence of which the distance between two consecutive repeaters (used for amplifying and reshaping the attenuated signals) could be as large as 250 km. We should perhaps mention here that it was the important paper of Kao and Hockham in 1966⁵ that suggested that optical fibers based on silica glass could provide the necessary transmission medium if metallic and other impurities could be removed. To quote Kao and Hockham:

Theoretical and experimental studies indicate that a clad glass fiber with a core diameter of about λ_0 and

**The attenuation is usually measured in dB (decibels) — we will define this in Sec. 24.8. We may mention that a loss of $0.25\ \text{dB/km}$ would imply that the power will decrease by a factor of 10 in traversing a distance of 40 km.

an overall diameter of about $1000 \lambda_0$ represents a possible practical optical waveguide with important potential as a new form of communication medium. The refractive index of the core needs to be about 1% higher than that of cladding. However, the attenuation should be around 20 dB/km which is much higher than the lower limit of loss figure imposed by fundamental mechanisms.

Indeed this 1966 paper triggered the beginning of serious research in developing low loss optical fibers. In 1970, Kapron, Keck and Maurer (at Corning Glass in USA) were successful in producing silica fibers with a loss of about 17 dB/km at a wavelength of 633 nm.⁶ Since then, the technology has advanced with tremendous rapidity. By 1985 glass fibers were routinely produced with extremely low losses (< 0.25 dB/km).

Figure 24.3 shows a typical optical fiber communication system. It consists of a transmitter which could be either a laser diode or an LED, the light from which is coupled into an optical fiber. Along the path of the optical fiber, there are splices which are permanent joints between sections of fi-

bers and also repeaters which boost the signal and correct any distortion that may have accumulated along the path of the fiber. At the end of the link, the light is detected by a photodetector and electronically processed to retrieve the signal.

24.3 TOTAL INTERNAL REFLECTION

At the heart of an optical communication system is the optical fiber that acts as the transmission channel carrying the light beam loaded with information; and as mentioned earlier, the guidance of the light beam (through the optical fiber) takes place because of the phenomenon of total internal reflection (often abbreviated as TIR). Now, if a ray is incident at the interface of a denser medium ($n_2 > n_1$), the refracted ray will bend towards the normal [see Fig. 24.4(a)]. On the other hand, if a ray is incident at the interface of a rarer medium ($n_2 < n_1$) then the ray will bend away from the normal [see Fig. 24.4(b)]. The angle of

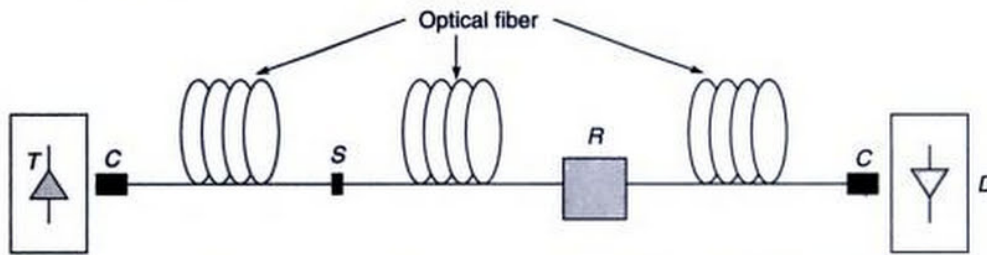


Fig. 24.3 Typical optical fiber communication system. It consists of a transmitter (shown as T) which could be either a laser diode or a LED, the light from which is coupled into an optical fiber. Along the path of the optical fiber, there are splices (shown as S) which are permanent joints between sections of fibers and also repeaters (shown as R) which boost the signal and correct any distortion that may have accumulated along the path of the fiber. At the end of the link, the light is detected by a photo detector (shown as D) and processed to retrieve the signal; C represents optical connectors.

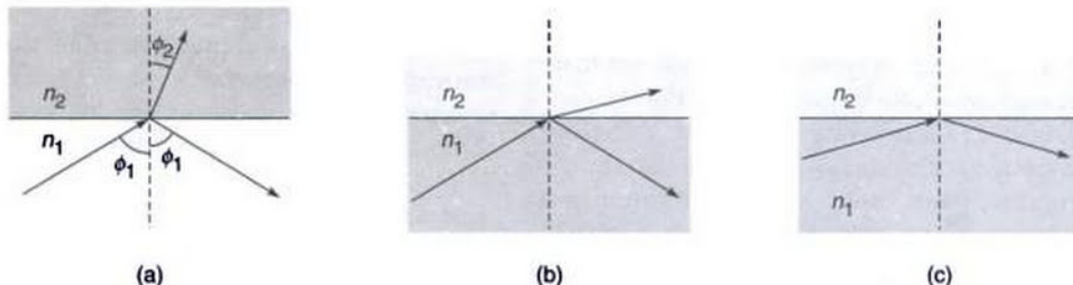


Fig. 24.4 (a) For a ray incident on a denser medium ($n_2 > n_1$), the angle of refraction is less than the angle of incidence, (b) For a ray incident on a rarer medium ($n_2 < n_1$), the angle of refraction is greater than the angle of incidence, (c) if the angle of incidence is greater than critical angle, it will undergo total internal reflection.

incidence, for which the angle of refraction is 90° , is known as the critical angle and is denoted by ϕ_c . Thus, when

$$\phi_1 = \phi_c = \sin^{-1} \left(\frac{n_1}{n_2} \right) \quad (1)$$

$\phi_2 = 90^\circ$. When the angle of incidence exceeds the critical angle (i.e., when $\phi_1 > \phi_c$), there is no refracted ray and we have what is known as total internal reflection (TIR)—see Fig. 24.4(b).

Example 24.1 For the glass–air interface, $n_1 = 1.5$, $n_2 = 1.0$ and the critical angle is given by

$$\phi_c = \sin^{-1} \left(\frac{1.0}{1.5} \right) \approx 41.8^\circ$$

On the other hand, for the glass–water interface, $n_1 = 1.5$, $n_2 = 4/3$ and

$$\phi_c = \sin^{-1} \left(\frac{1.33}{1.5} \right) \approx 62.7^\circ$$

The phenomenon of total internal reflection can be very easily demonstrated through a simple experiment as shown in Fig. 24.5. A thick semi-circular glass disc is immersed in a glass vessel filled with water. A laser beam from a He–Ne laser or simply a laser pointer is directed towards the centre of the semi-circular disc so that it is incident normally on the glass surface and goes undeviated as shown in the figure. The angle of incidence (at the glass–water interface) is increased by rotating the glass disc about the point O ; eventually when the angle of incidence exceeds the critical angle ($\approx 62.7^\circ$), the laser beam undergoes total internal reflection which can be clearly seen when viewed from the top. If one puts in a drop of ink in water, the light path becomes very beautiful to look at! The experiment is very simple and we urge the reader to carry it out using a laser pointer.

The first experimental demonstration of light guidance using total internal reflection was carried out by sending a light beam in a water jet; this was first demonstrated by

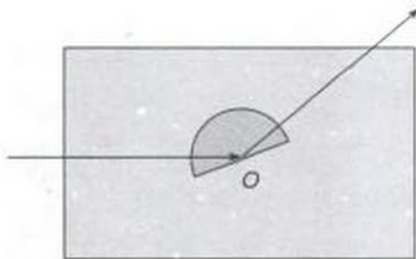


Fig. 24.5 A simple laboratory experiment to demonstrate the phenomenon of total internal reflection.

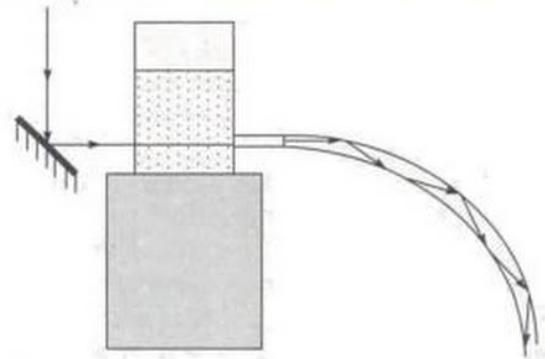


Fig. 24.6 Light guidance through a water jet using the phenomenon of total internal reflection; this was first demonstrated by Daniel Colladon in 1841.

Daniel Colladon in 1841 and also by Jacques Babinet in 1842. A schematic of this demonstration is shown in Fig. 24.6; light undergoes total internal reflection at the water–air interface and travels along the curved path of water emanating from an illuminated vessel. We should mention here that John Tyndall is usually credited with the first demonstration of light guidance in water jets; however, he demonstrated light guiding in water jets only in 1854, duplicating but not acknowledging Babinet.*

24.4 THE OPTICAL FIBER

Figure 24.7(a) shows an optical fiber, which consists of a (cylindrical) central dielectric core cladded by a material of slightly lower refractive index. The corresponding refractive index distribution (in the transverse direction) is given by

$$\begin{aligned} n &= n_1 \quad r < a \\ &= n_2 \quad r > a \end{aligned} \quad (2)$$

where n_1 and n_2 ($< n_1$) represent respectively the refractive indices of core and cladding and a represents the radius of the core. We define a parameter Δ through the following equation:

$$\Delta \equiv \frac{n_1^2 - n_2^2}{2n_1^2} \quad (3)$$

When $n_1 \approx n_2$, i.e., when $\Delta \ll 1$ (as is true for most silica fibers)

$$\begin{aligned} \Delta &= \frac{n_1 - n_2}{n_1} \frac{n_1 + n_2}{2n_1} \\ &\approx \frac{n_1 - n_2}{n_2} \approx \frac{n_1 - n_2}{n_1} \end{aligned} \quad (4)$$

*For a very nice historical survey, we refer the reader to Ref. 1.

For a typical (multimoded) fiber, $a \approx 25 \mu\text{m}$, $n_2 \approx 1.45$ (pure silica) and $\Delta \approx 0.01$ giving a core index of $n_1 \approx 1.465$. The cladding is usually pure silica while the core is usually silica doped with germanium; doping by germanium results in an increase of refractive index.

Now, for a ray entering the fiber, if the angle of incidence (at the core-cladding interface) is greater than the critical angle ϕ_c , then the ray will undergo TIR at that interface. Further, because of the cylindrical symmetry in the fiber structure, this ray will suffer TIR at the lower interface also and therefore get guided through the core by repeated total internal reflections. Even for a bent fiber, light guidance can occur through multiple total internal reflections (see Figures 24.2 and 24.6). Figure 24.2 shows the actual guidance of a light beam as it propagates through a long optical fiber; in the photograph, the light emerging from the side of the fiber is due mainly to Rayleigh scattering, the same phenomenon that is responsible for the blue color of the sky and the red color of the rising or the setting sun.

The necessity of a cladded fiber (Fig. 24.7) rather than a bare fiber, i.e. without a cladding, was felt because of the fact that for transmission of light from one place to another, the fiber must be supported, and supporting structures may considerably distort the fiber thereby affecting the guidance of the light wave. This can be avoided by choosing a sufficiently thick cladding. Further, in a fiber bundle, in the

absence of the cladding, light can leak through from one fiber to another. The idea of adding a second layer of glass (namely, the cladding) came in 1955 simultaneously from van Heel in Holland and from Hopkins and Kapany in the UK; however, during that time the use of optical fibers was mainly in image transmission rather than in communications. Indeed, the early pioneering works in fiber-optics (in the 1950's) were by Hopkins and Kapany and by Van Heel; these works led to the use of the fiber in optical devices.

It is of interest to mention that the retina of the human eye consists of a large number of rods and cones which have the same kind of structure as the optical fiber, i.e. they consist of dielectric cylindrical rods surrounded by another dielectric of slightly lower refractive index (see Fig. 24.8). The core diameters are in the range of a few microns. The light absorbed in these "light guides" generates electrical signals, which are then transmitted to the brain through various nerves.

24.5 WHY GLASS FIBERS?

Why are optical fibers made of glass? Quoting Professor W.A. Gambling, who is one of the pioneers in the field of fiber-optics:²

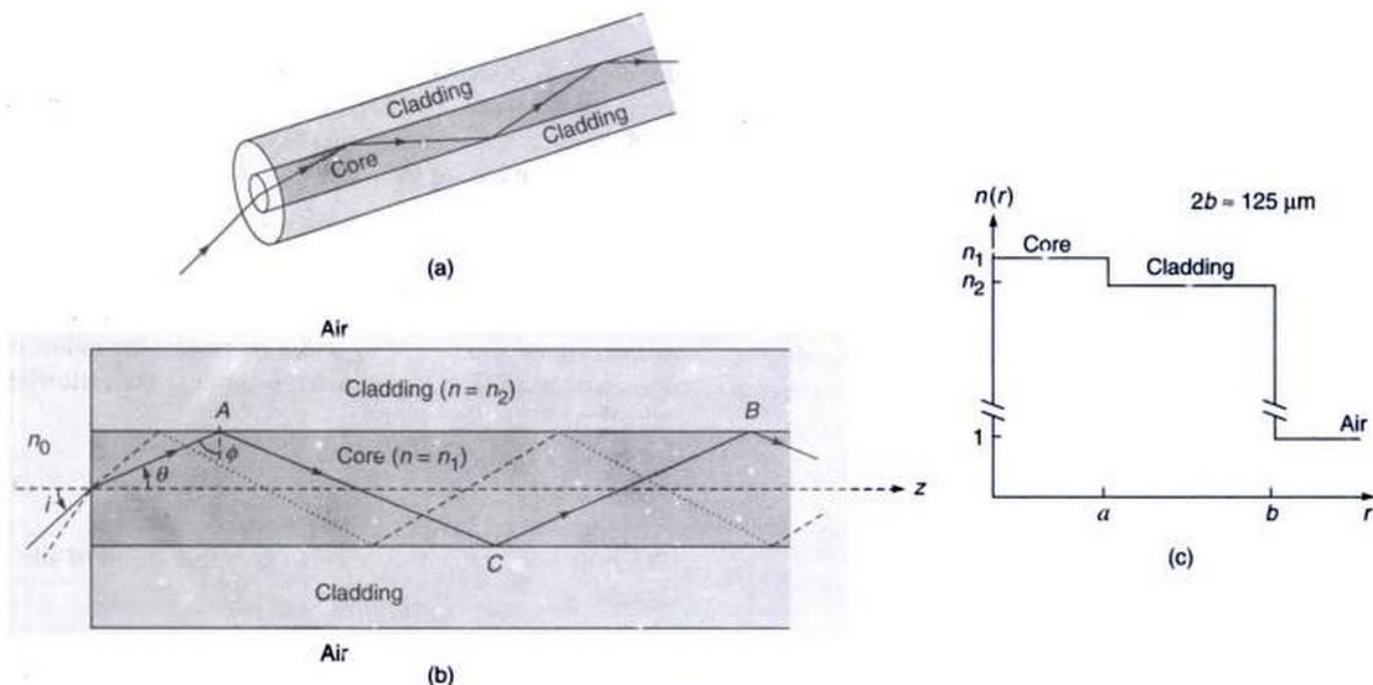


Fig. 24.7 (a) A glass fiber consists of a cylindrical central core cladded by a material of slightly lower refractive index. (b) Light rays incident on the core-cladding interface at an angle greater than the critical angle are trapped inside the core of the fiber. (c) Refractive index distribution for a step-index fiber.

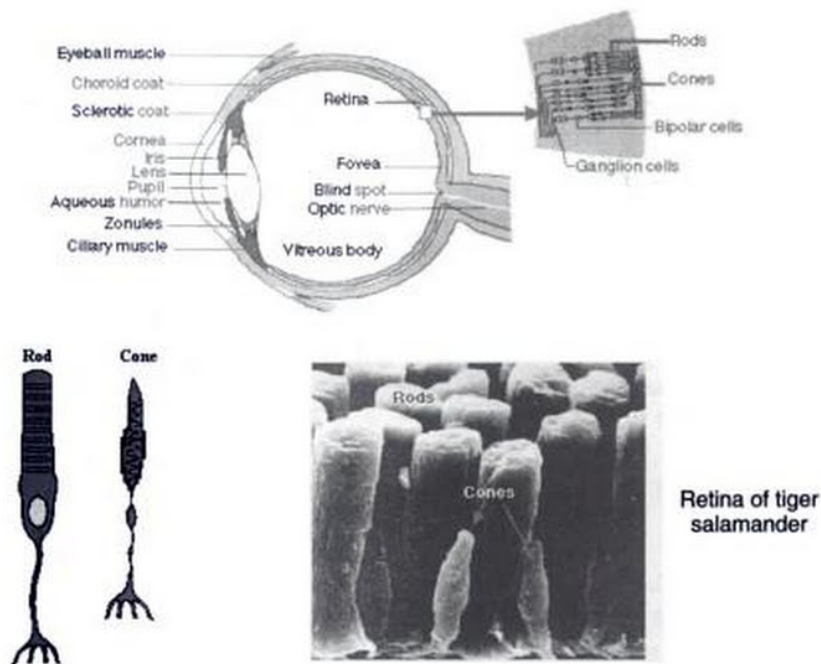


Fig. 24.8 The rods and cones of the eye [Photograph downloaded from the internet by Professor Thyagarajan].

We note that glass is a remarkable material which has been in use in 'pure' form for at least 9000 years. The compositions remained relatively unchanged for millennia and its uses have been widespread.

The three most important properties of glass which makes it of unprecedented value are:

1. First, there is a wide range of accessible temperatures where its viscosity is variable and can be well controlled unlike most materials, like water and metals which remain liquid until they are cooled down to their freezing temperatures and then suddenly become solid. Glass, on the other hand, does not solidify at a discrete freezing temperature but gradually becomes stiffer and stiffer and eventually becoming hard. In the transition region it can be easily drawn into a thin fiber.
2. The second most important property is that highly pure silica is characterized with extremely low-loss; i.e., it is highly transparent. Today in most commercially available silica fibers, 96 % of the power gets transmitted after propagating through 1 km of optical fiber. This indeed represents a truly remarkable achievement.
3. The third most remarkable property is the intrinsic strength of glass. Its strength is about $2,000,000 \text{ lb/in}^2$ so that a glass fiber of the type used in the tele-

phone network and having a diameter ($125 \mu\text{m}$) of twice the thickness of a human hair, can support a load of 40 lb.

24.6 THE COHERENT BUNDLE

If a large number of fibers are put together, it forms what is known as a bundle. If the fibers are not aligned, i.e., they are all jumbled up, the bundle is said to form an incoherent bundle. However, if the fibers are aligned properly, i.e., if the relative positions of the fibers in the input and output ends are the same, the bundle is said to form a coherent bundle. Now, if a particular fiber is illuminated at one of its ends, then there will be a bright spot at the other end of the same fiber; thus a coherent bundle will transmit the image from one end to another (see Fig. 24.9). On the other hand, in an incoherent bundle the output image will be scrambled. Because of this property, an incoherent bundle can be used as a coder, the transmitted image can be decoded by using a similar bundle at the output end. In a bundle, since there can be hundreds of thousands of fibers, decoding without the original bundle configuration should be extremely difficult. Incoherent bundles are also used in illumination such as in traffic lights or road signs* or even for lighting in real estate applications wherein the light source is removed from relatively inaccessible areas and fibers are used to

*[see e.g., Ref. 7]

defines the numerical aperture (NA) of the fiber by the following equation:

$$\text{NA} = \sin i_m = \sqrt{n_1^2 - n_2^2} \quad (6)$$

Example 24.2 For a typical step-index (multimode) fiber with $n_1 \approx 1.45$ and $\Delta \approx 0.01$, we get

$$\sin i_m \approx 0.205 \Rightarrow i_m \approx 12^\circ$$

Now, in a short length of an optical fiber, if all rays between $i = 0$ and i_m are launched, then, the light coming out of the fiber will also appear as a cone of semi-angle i_m emanating from the fiber end. If we now allow this beam to fall normally on a white paper (see Fig. 24.11) and measure its diameter we can easily calculate the NA of the fiber. This allows us to estimate the NA of the optical fiber by a very simple experiment. The procedure is as follows:

Several concentric circles of increasing radii—say, starting from 0.5 cm to 1.5 cm, are drawn on a small paper screen and the screen is positioned in the far-field such that the axis of the fiber, at the output end, passes perpendicularly through the center of these circles on the screen. The fiber end, which is mounted on a XYZ-stack, is moved slightly towards or away from the screen so that one of the circles just circumscribes the far-field radiation spot. The distance z between the fiber-end and the screen, and the diameter D of the coinciding circle are measured accurately. The NA is calculated using the following equation

$$\text{NA} = \sin i_m = \sin \left[\tan^{-1} \left(\frac{D}{2z} \right) \right] \quad (7)$$

24.8 ATTENUATION IN OPTICAL FIBERS

Attenuation and pulse dispersion represents the two most important characteristics of an optical fiber that determine

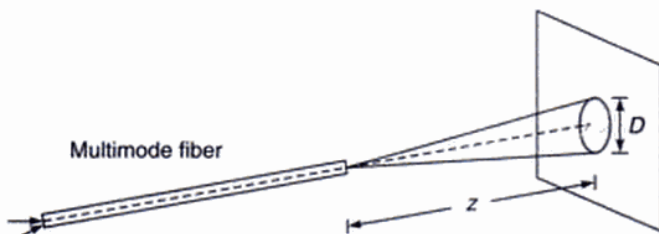


Fig. 24.11 Measurement of the diameter D of the spot on a screen placed at a far-field distance z from the output end of a multimode fiber can be used to measure the NA of the fiber.

the information carrying capacity of a fiber-optic communication system. Obviously, lower the attenuation (and similarly lower the dispersion) greater will be the required repeater spacing and therefore lower will be the cost of the communication system. Pulse dispersion will be discussed in Sec. 24.10 while in this section we will briefly discuss the various attenuation mechanisms in an optical fiber.

The attenuation of an optical beam is usually measured in decibels (dB). If an input power P_1 results in an output power P_2 , then the loss in decibels is given by

$$\alpha = 10 \log_{10} \left(\frac{P_{\text{input}}}{P_{\text{output}}} \right) \quad (8)$$

Thus if the output power is only half of the input power, then the loss is $10 \log 2 \approx 3$ dB. Similarly, 20 dB and 10 dB loss will correspond to power reduction by a factor of 100 and 10 respectively. On the other hand, if 96% of the light is transmitted through the fiber, the loss is about 0.18 dB. In a typical fiber amplifier a power amplification by a factor of 1000 implies a power gain of 30 dB.

Figure 24.12 shows the spectral dependence of loss coefficient (i.e., loss per unit length) of a typical silica optical fiber. One can notice two important low loss windows around $1.3 \mu\text{m}$ and $1.55 \mu\text{m}$ (both wavelengths are actually in the infrared region). The typical losses at these wavelengths are 0.3–0.4 dB/km and about 0.25 dB/km respectively. This is the reason why most fiber-optic systems operate either in the $1.3 \mu\text{m}$ window or $1.55 \mu\text{m}$ window. The latter window has become extremely important in view of the availability of optical amplifiers (see Sec. 23.1).

The losses are caused due to various mechanisms such as Rayleigh scattering, absorption due to metallic impurities and water and due to intrinsic absorption of silica molecule itself. Even 1 ppm (part per million) of iron can cause a loss

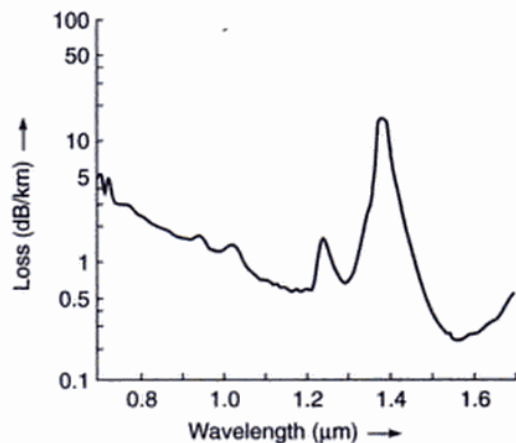


Fig. 24.12 Typical wavelength dependence of loss for a silica fiber. Notice that the lowest loss occurs at 1550 nm [adapted from Ref. 9].

the fiber is said to be a single-mode fiber. However, if

$$V \geq 10$$

the fiber is said to be a multimoded fiber. Now, while discussing step-index fibers, we had considered light propagation inside the fiber as a set of many rays bouncing back and forth at the core-cladding interface (see Fig. 24.7) and the angle θ could take *all* possible values from 0 (corresponding to a ray propagating parallel to the z -axis) to $\cos^{-1}(n_2/n_1)$ (corresponding to a ray incident at the critical angle on the core-cladding interface). However, if we do carry out an experiment of the type shown in Fig. 24.13, we will find that the output beam corresponds to only discrete values of θ . Each 'discrete' ray path correspond to what is known as the mode of the waveguide. Indeed, in the experiment shown in Fig. 24.13, if we measure the discrete values of θ , we can obtain the discrete values of the propagation constant β by using the following formula:

$$\beta = k_0 n_1 \cos \theta$$

Now, if $V < 2.4045$, we will have only one discrete ray path and the fiber is said to be a single-mode fiber. For such a fiber, one should use wave optics to study the propagation characteristics of the optical fiber and ray optics will not be applicable. Further, for a single-mode fiber there is only one guided mode and the transverse field distribution $\psi(x, y)$ associated with the mode is approximately Gaussian—we will discuss this in Sec. 24.9.1.

On the other hand, for $V \geq 10$, the number of modes is approximately given by $\frac{1}{2}V^2$ and the fiber is said to be a multimoded fiber. Different modes (in a multimoded fiber) travel with different group velocities leading to what is known as intermodal dispersion; in the language of ray optics, this is known as ray dispersion arising due to the fact that different rays take different amounts of time in propagating through the fiber (see Sections 24.2, 24.10.1 and 24.10.2). Indeed in a highly multimoded fiber, we can use ray optics to calculate pulse dispersion.

Example 24.6 Consider a step-index fiber (operating at $1.3 \mu\text{m}$) with $n_2 = 1.447$, $\Delta = 0.003$ and $a = 4.2 \mu\text{m}$.

$$\text{Now, } V = \frac{2\pi}{1.3} \times 4.2 \times 1.447 \times \sqrt{0.006} \approx 2.275$$

Thus the fiber is single moded at $1.3 \mu\text{m}$. At any other wavelength

$$V = \frac{2.958}{\lambda_0}$$

where λ_0 is measured in μm . Thus for

$$\lambda_0 > 2.958/2.4045 = 1.23 \mu\text{m}$$

the fiber will be single moded. The wavelength for which $V = 2.4045$ is known as the *cut-off wavelength* and is denoted by λ_c . In this example, $\lambda_c = 1.23 \mu\text{m}$ and the fiber will be single moded for $\lambda_0 > 1.23 \mu\text{m}$.

Example 24.7 For reasons that will be discussed later, the fibers used in IV generation optical communication systems (operating at $1.55 \mu\text{m}$) have a small value of core radius and a large value of Δ . A typical fiber (operating at $\lambda_0 \approx 1.55 \mu\text{m}$) would have $n_1 = 1.46$, $\Delta = 0.0075$ and $a = 2.3 \mu\text{m}$. Thus at $\lambda_0 = 1.55 \mu\text{m}$

$$V = \frac{2\pi}{1.55} \times 2.3 \times 1.46 \times \sqrt{0.015} \approx 1.667$$

The fiber will be single moded (at $1.55 \mu\text{m}$). Further, for the given fiber we may write

$$V = \frac{2.584}{\lambda_0}$$

and therefore the cut off wavelength will be

$$\lambda_c = 2.584/2.4045 = 1.07 \mu\text{m}.$$

24.9.1 Spot Size of the Fundamental Mode

As mentioned earlier, a single-mode fiber supports only one mode that propagates through the fiber; this is also referred

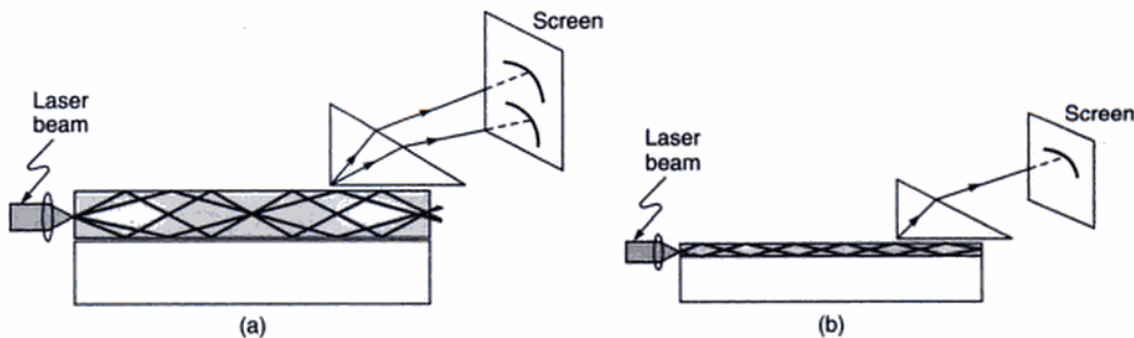


Fig. 24.13 A laser beam is focused at the entrance aperture of a waveguide. (a) Each 'discrete' ray path correspond to a mode of the waveguide. (b) For a single-mode waveguide there is only one discrete value of θ .

to as the fundamental mode of the fiber. The transverse field distribution associated with the fundamental mode of a single-mode fiber is an extremely important quantity and it determines various important parameters like splice loss at joints, launching efficiencies, bending loss etc. For most single-mode fibers, the fundamental mode field distributions can be well approximated by a Gaussian function, which may be written in the form

$$\psi(x, y) = A e^{-\frac{x^2+y^2}{w^2}} = A e^{-\frac{r^2}{w^2}} \quad (16)$$

where w is referred to as the spot size of the mode field pattern. For a step-index single-mode fiber one has the following empirical expression for w ¹¹:

$$\frac{w}{a} \approx 0.65 + \frac{1.619}{V^{3/2}} + \frac{2.879}{V^6}; \quad 0.8 \leq V \leq 2.5 \quad (17)$$

where a is the core radius. We may mention here that the light coming out of a He-Ne laser (or of a laser pointer) has a transverse intensity distribution very similar to that coming out from a single-mode fiber except that the spot size is much larger.

Example 24.8 Consider a step-index fiber (operating at 1300 nm) with $n_2 = 1.447$, $\Delta = 0.003$ and $a = 4.2 \mu\text{m}$ (see Example 24.6). Thus $V \approx 2.28$ giving $w = 4.8 \mu\text{m}$. The same fiber will have a V value of 1.908 at $\lambda_0 = 1550 \text{ nm}$ giving a value of the spot size $\approx 5.5 \mu\text{m}$. Thus the spot size increases with wavelength.

Example 24.9 For a step-index fiber (operating at 1550 nm) with $n_1 = 1.46$, $\Delta = 0.0075$ and $a = 2.3 \mu\text{m}$ (see Example 24.7). Thus $V \approx 1.67$ giving $w \approx 3.5 \mu\text{m}$. The same fiber will have a V value of 1.99 at $\lambda_0 = 1300 \text{ nm}$ giving a value of the spot size $\approx 2.9 \mu\text{m}$.

24.9.2 Splice Loss Due to Transverse Misalignment

The most common misalignment at a joint between two similar fibers is the transverse misalignment similar to that shown in Fig. 24.25. Corresponding to a transverse misalignment of u the loss in decibels is given by

$$\alpha \text{ (dB)} \approx 4.34 (u/w)^2 \quad (18)$$

Thus a larger value of w will lead to a greater tolerance to transverse misalignment. For $w \approx 5 \mu\text{m}$, and a transverse offset of $1 \mu\text{m}$ the loss at the joint will be approximately 0.18 dB; on the other hand, for $w \approx 3 \mu\text{m}$, a transverse offset of $1 \mu\text{m}$ will result in a loss of about 0.5 dB.

Example 24.10 For a single-mode fiber operating at 1300 nm, $w = 5 \mu\text{m}$, and if the splice loss is to be below 0.1 dB, then from Eq. (18) we obtain $u < 0.76 \mu\text{m}$. Thus, for a

low-loss joint, the transverse alignment is very critical and connectors for single-mode fibers require precision matching and positioning for achieving low loss.

Typical transverse cross sections and corresponding ray paths in different types of optical fibers are shown in Figs 24.14 and 24.15.

24.10 PULSE DISPERSION IN OPTICAL FIBERS

In digital communication systems, information to be sent is first coded in the form of pulses and then these pulses of

Transverse cross sections of different types of optical fibers

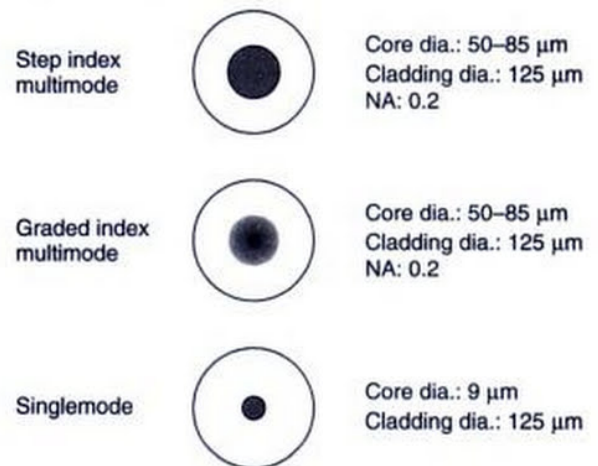


Fig. 24.14 Transverse cross sections and typical dimensions of step index multimode, graded index multimode and single mode fibers.

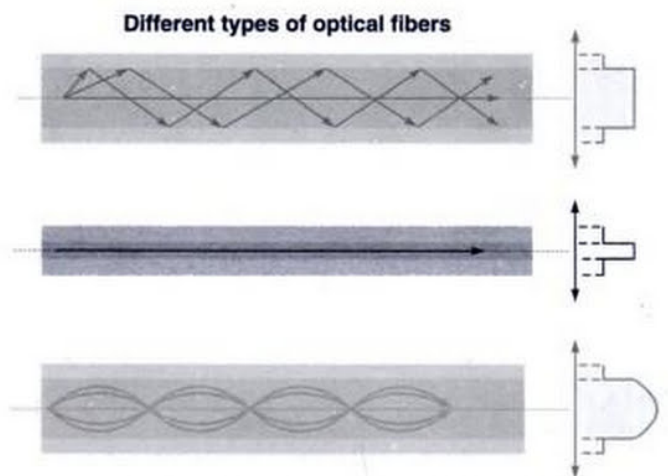


Fig. 24.15 Ray paths in step index multimode, single mode and graded index multimode optical fibers.

light are transmitted from the transmitter to the receiver where the information is decoded. Larger the number of pulses that can be sent per unit time and still be resolvable at the receiver end, larger would be the transmission capacity of the system. A pulse of light sent into a fiber broadens in time as it propagates through the fiber; this phenomenon is known as pulse dispersion and occurs primarily because of the following mechanisms:

1. In multimode fibers, different rays take different times to propagate through a given length of the fiber; we will discuss this for a step-index fiber and for a parabolic index fiber in this and the following sections. In the language of wave optics, this is known as *intermodal dispersion* because it arises due to different modes travelling with different group velocities.
2. Any given light source emits over a range of wavelengths and, because of the intrinsic property of the material of the fiber, different wavelengths take different amounts of time to propagate along the same path. This is known as material dispersion and obviously, it is present in both single-mode and multimode fibers.
3. On the other hand, in single-mode fibers since there is only mode, there is no intermodal dispersion. However, we have what is known as intramodal dispersion. Physically, this arises due to the fact that the spot size (of the fundamental mode) itself depends on the wavelength [see Examples 24.8 and 24.9]. Obviously, intramodal dispersion is present in multimode fibers also—however, the effect is very small and can be neglected.

24.10.1 Ray Dispersion in Step-Index Fibers

We first consider ray paths in a step-index fiber as shown in Fig. 24.7. As can be seen, rays making larger angles with the axis (those shown as dotted rays) have to traverse a longer optical path length and therefore take a longer time to reach the output end.

We will derive an expression for the intermodal dispersion for a step-index fiber. Referring back to Fig. 24.7, for a ray making an angle θ with the axis, the distance AB is traversed in time

$$t_{AB} = \frac{AC + CB}{c/n_1} = \frac{AB/\cos \theta}{c/n_1} \quad (19)$$

or

$$t_{AB} = \frac{n_1 AB}{c \cos \theta} \quad (20)$$

where c/n_1 represents the speed of light in a medium of refractive index n_1 , c being the speed of light in free space.

Since the ray path will repeat itself, the time taken by a ray to traverse a length L of the fiber would be

$$t_L = \frac{n_1 L}{c \cos \theta} \quad (21)$$

The above expression shows that the time taken by a ray is a function of the angle θ made by the ray with the z -axis, which leads to pulse dispersion. If we assume that all rays lying between $\theta = 0$ and $\theta = \theta_c = \cos^{-1}(n_2/n_1)$ [see Eq. (5)] are present, then the time taken by these extreme rays for a fiber of length L would be given by

$$t_{\min} = \frac{n_1 L}{c} \text{ corresponding to } \theta = 0 \quad (22)$$

$$t_{\max} = \frac{n_1 L}{c n_2} \text{ corresponding to } \theta = \theta_c = \cos^{-1}(n_2/n_1) \quad (23)$$

Hence if all the input rays were excited simultaneously, the rays would occupy a time interval at the output end of duration

$$\Delta \tau_i = t_{\max} - t_{\min} = \frac{n_1 L}{c} \left[\left(\frac{n_1}{n_2} \right) - 1 \right]$$

or

$$\Delta \tau_i \cong \frac{n_1 L}{c} \Delta \approx \frac{L}{2n_1 c} (\text{NA})^2$$

Intermodal dispersion in multimode SIF (24)

where Δ has been defined earlier [see Eqs (3) and (4)] and we have used Eq. (6). The quantity $\Delta \tau_i$ represents the pulse dispersion due to different rays taking different times in propagating through the fiber which, in wave optics, is nothing but the intermodal dispersion and hence the subscript i . Note that the pulse dispersion is proportional to the square of NA. Thus to have a smaller dispersion, one must have a smaller NA which of course reduces the acceptance angle and hence the light gathering power. Now, if at the input end of the fiber, we have a pulse of width τ_1 then after propagating through a length L of the fiber, the pulse would have a width τ_2 given approximately by

$$\tau_2^2 = \tau_1^2 + \Delta \tau_i^2 \quad (25)$$

Consequently, the pulse broadens as it propagates through the fiber (see Fig. 24.16). Hence, even though two pulses may be well resolved at the input end, because of the broadening of the pulses they may not be so at the output end.

Example 24.11 For a typical (multimoded) step-index fiber, if we assume $n_1 = 1.5$, $\Delta = 0.01$, $L = 1$ km, we would get

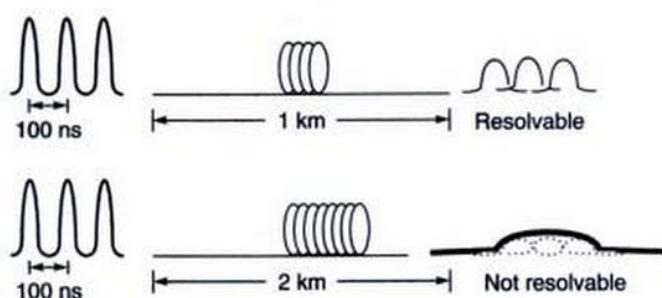


Fig. 24.16 Pulses separated by 100 ns are incident on a fiber characterized by a dispersion of 50 ns/km. They would be resolvable at the output end of 1 km of the fiber. The same pulses would not be resolvable at the output end of 2 km of the same fiber.

$$\Delta\tau_i = \frac{1.5 \times 1000}{3 \times 10^8} \times 0.01 = 50 \text{ ns/km} \quad (26)$$

i.e. a pulse after traversing through the fiber of length 1 km will be broadened by 50 ns. Thus two pulses separated by, say, 500 ns at the input end would be quite resolvable at the end of 1 km of the fiber. However, if consecutive pulses are separated by, say, 10 ns at the input end, they would be absolutely unresolvable at the output end. Hence in a 1 Mbit/s fiber-optic system, where we have one pulse every 10^{-6} s, a 50 ns/km dispersion would require repeaters to be placed every 3–4 km. On the other hand, in a 1 Gbit/s fiber-optic communication system, which requires the transmission of one pulse every 10^{-9} s, a dispersion of 50 ns/km would result in intolerable broadening even within 50 metres or so which would be highly inefficient and uneconomical from a system point of view.

Where the output pulses are not resolvable, no information can be retrieved. Thus, smaller the pulse dispersion, greater will be the information carrying capacity of the system.

From the discussion in the above example it follows that for a very high information carrying system, it is necessary to reduce the pulse dispersion; two alternative solutions exist—one involves the use of near parabolic index fibers and the other involves single-mode fibers.

24.10.2 Parabolic-Index Fibers (PIF)

In a step-index fiber such as that shown in Fig. 24.7, the refractive index of the core has a constant value. By contrast, in a parabolic index fiber, the refractive index in the core decreases continuously (in a quadratic fashion) from a maximum value at the center of the core to a constant value at the core-cladding interface. The refractive index variation is given by

$$\begin{aligned} n^2(r) &= n_1^2 \left[1 - 2\Delta \left(\frac{r}{a} \right)^2 \right] & 0 < r < a & \text{core} \\ &= n_2^2 = n_1^2 (1 - 2\Delta) & r > a & \text{cladding} \end{aligned} \quad (27)$$

with Δ as defined in Eq. (4). For a typical (multimode) parabolic index silica fiber $\Delta \approx 0.01$, $n_2 \approx 1.45$ and $a \approx 25 \mu\text{m}$. In Sec. 2.4.1 we had shown that the ray paths in a parabolic waveguide are sinusoidal (see Fig. 24.17 which is the same as Fig. 2.17).

Now, even though rays making larger angles with the axis traverse a larger path length, they do so now in a region of lower refractive index (and hence greater speed). The longer path length is almost compensated for by a greater average speed such that all rays take approximately the same amount of time in traversing the fiber. In Sec. 2.4.2 we had made a detailed calculation of determining the time taken by a particular ray to propagate through a parabolic index waveguide; the final result for the intermodal dispersion is given by:

$$\Delta\tau_i = \frac{n_2 L}{2c} \left(\frac{n_1 - n_2}{n_2} \right)^2 \approx \frac{n_2 L}{2c} \Delta^2 \approx \frac{L}{8cn_1^3} (\text{NA})^4 \quad (28)$$

Note that as compared to a step-index fiber, the pulse dispersion is proportional to the fourth power of NA. For a typical (multimode parabolic index) fiber with $n_2 \approx 1.45$ and $\Delta \approx 0.01$, we would get

$$\Delta\tau_i \approx 0.25 \text{ ns/km} \quad (29)$$

Comparing with Eq. (26) we find that for a parabolic index fiber the pulse dispersion is reduced by a factor of about 200 in comparison to the step-index fiber. It is because of this reason that first and second generation optical communication systems used near parabolic index fibers. In order to further decrease the pulse dispersion, it is necessary

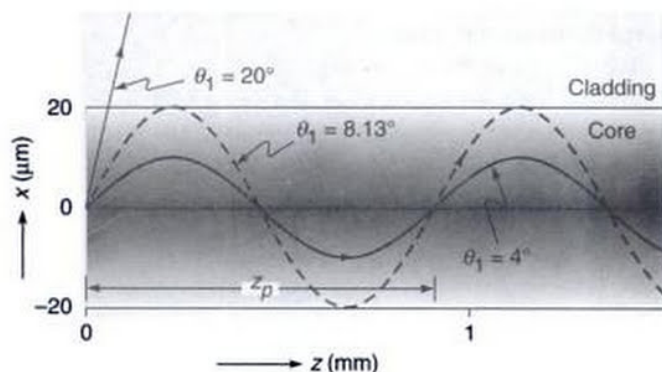


Fig. 24.17 Ray paths in a parabolic index fiber.

to use single-mode fibers because there will be no intermodal dispersion. However, in all fiber-optic systems we will have material dispersion which is a characteristic of the material itself and not of the waveguide; we will discuss this in the following section.

24.10.3 Material Dispersion

We have considered the broadening of an optical pulse due to different rays taking different amounts of time to propagate through a certain length of the fiber. However, every source of light has a certain wavelength spread which is often referred to as the *spectral width of the source*. Thus a white light source (like the sun) would have a spectral width of about 300 nm; on the other hand, an LED would have a spectral width of about 25 nm and a typical laser diode (LD) operating at 1300 nm would have a spectral width of about 2 nm or less. In Chapter 8 we had discussed that the refractive index of the medium (and hence the group velocity v_g) depends on the wavelength. Thus, each wavelength component (of the pulse) will travel with a slightly different group velocity through the fiber, resulting in a broadening of a pulse. In Sec. 8.2 we had shown that the pulse broadening (due to wavelength dependence of the refractive index) is given by

$$\Delta\tau_m = -\frac{L\Delta\lambda_0}{\lambda_0 c} \left[\lambda_0^2 \frac{d^2 n}{d\lambda_0^2} \right] \quad (30)$$

where L is the length of the fiber, $\Delta\lambda_0$ is the spectral width of the source and c the speed of light in free space; the subscript m in Eq.(30) refers to the fact that we are considering material dispersion. We had also defined the material dispersion coefficient (which is measured in ps/km-nm):

$$D_m = \frac{\Delta\tau_m}{L\Delta\lambda_0} = -\frac{10^4}{3\lambda_0} \left[\lambda_0^2 \frac{d^2 n}{d\lambda_0^2} \right] \text{ ps/km.nm} \quad (31)$$

where we have used $c \approx 3 \times 10^7$ km/ps and λ_0 is measured in μm and the quantity inside the square brackets is dimensionless. Thus D_m represents the material dispersion in picoseconds per kilometer length of the fiber per nanometer spectral width of the source. At a particular wavelength, the value of D_m is a characteristic of the material and is (almost) the same for *all* silica fibers. The values of D_m for different wavelengths (for pure silica) are tabulated in Table 8.1. When D_m is negative, it implies that the longer wavelengths travel faster; similarly, a positive value of D_m implies that shorter wavelengths travel faster. [see Fig. 8.3].

Example 24.12 The LEDs used in the earlier optical communication systems had a spectral width $\Delta\lambda_0$ of about 20 nm around $\lambda_0 = 825$ nm; at this wavelength $D_m \approx 84.2$ ps/km.nm. Thus a pulse will broaden by

$$\begin{aligned} \Delta\tau_m &= D_m \times L \times \Delta\lambda \\ &= 84.2 \text{ (ps/km-nm)} \times 1 \text{ (km)} \times 20 \text{ (nm)} \\ &\approx 1700 \text{ ps} = 1.7 \text{ ns} \end{aligned}$$

in traversing 1 km length of the fiber. It is interesting to note that if we carry out a similar calculation around $\lambda_0 \approx 1300$ nm (where $D_m \approx 2.4$ ps/km.nm), we will obtain a much smaller value of $\Delta\tau_m$; thus

$$\begin{aligned} \Delta\tau_m &= D_m \times L \times \Delta\lambda \\ &= 2.4 \text{ (ps/km-nm)} \times 1 \text{ (km)} \times 20 \text{ (nm)} \\ &\approx 0.05 \text{ ns} \end{aligned}$$

in traversing 1 km length of the fiber. The very small value of $\Delta\tau_m$ is due to the fact that v_g is approximately constant around $\lambda_0 = 1300$ nm, as shown in Fig. 8.3. Indeed the wavelength $\lambda_0 \approx 1270$ nm is usually referred to as the zero material dispersion wavelength, and it is because of such low material dispersion that the optical communication systems shifted their operation to around $\lambda_0 \approx 1300$ nm.

Example 24.13 In the optical communication systems that are in operation today, one uses LD's (Laser Diodes) with $\lambda_0 \approx 1550$ nm having a spectral width of about 2 nm; at this wavelength $D_m \approx 21.5$ ps/km - nm (see Table 8.1). Thus for a 1 km length of the fiber, the material dispersion $\Delta\tau_m$ becomes

$$\begin{aligned} \Delta\tau_m &= D_m \times L \times \Delta\lambda \\ &= 21.5 \text{ (ps/km-nm)} \times 1 \text{ (km)} \times 2 \text{ (nm)} \\ &\approx 43 \text{ ps} \end{aligned}$$

the positive sign indicating that higher wavelengths travel slower than lower wavelengths. [Notice from Table 8.1 that for $\lambda_0 \geq 1300$ nm, n_g increases with λ_0].

24.11 DISPERSION AND MAXIMUM BIT RATES

We may mention here briefly that in a digital communication system employing light pulses, pulse broadening would result in an overlap of pulses, resulting in loss of resolution and leading to errors in detection. Thus pulse broadening is one of the mechanisms (other than attenuation) that limits the distance between two repeaters in a fiber-optic link. It is obvious that larger the pulse broadening, the smaller would be the number of pulses per second that can be sent down a link. There are different criteria based on slightly different considerations that are used to estimate the maximum permissible bit rate (B_{\max}) for a given pulse dispersion. However, it is always of the order of $1/\Delta\tau$. In one type of extensively used coding [known as NRZ (Non Return to Zero)]

of each mode *does* depend on the wavelength; physically, this is due to the fact that the spot-size of the mode depends on the wavelength (see Examples 24.8 and 24.9)—this leads to what is known as *waveguide dispersion*. The detailed theory is rather involved* but a convenient empirical formula for a step-index single-mode fiber is given by

$$\Delta\tau_w = -\frac{L}{c} n_2 \Delta [0.080 + 0.549(2.834 - V)^2] \frac{\Delta\lambda_0}{\lambda_0}$$

$$\text{for } 1.4 < V < 2.6 \quad (34)$$

If we assume $L = 1 \text{ km} = 10^3 \text{ m}$, $\Delta\lambda_0 = 1 \text{ nm}$ and $c = 3 \times 10^8 \text{ m/ps}$ we get

$$D_w = -\frac{n_2 \Delta}{3\lambda_0} \times 10^7 [0.080 + 0.549(2.834 - V)^2] \text{ ps/km.nm} \quad (35)$$

where λ_0 is measured in nanometers. As before, the negative sign indicates that longer wavelengths travel faster. The total dispersion is given by the sum of material and waveguide dispersions:

$$D_{\text{tot}} = D_m + D_w \quad (36)$$

Let us consider the two single-mode fibers discussed earlier.

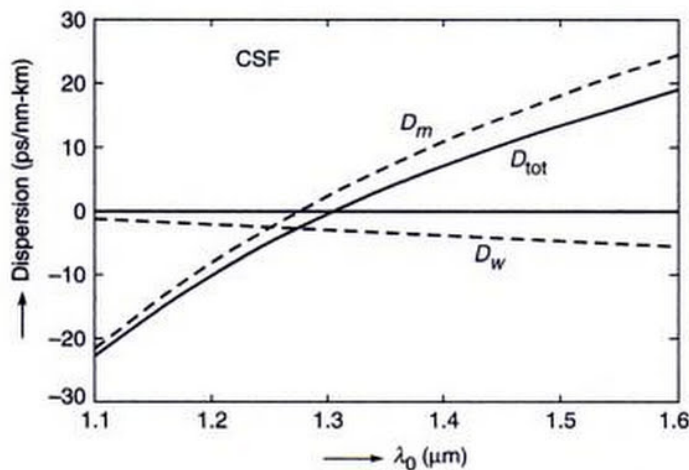


Fig. 24.18 The wavelength dependence of D_m , D_w and D_{tot} for a typical conventional single-mode fiber (CSF) with parameters as given in Example 24.6. The total dispersion passes through zero around $\lambda_0 \approx 1300 \text{ nm}$ which is known as zero dispersion wavelength.

Example 24.17 We consider the fiber discussed in Example 24.6 for which $n_2 = 1.447$, $\Delta = 0.003$ and $a = 4.2 \mu\text{m}$ so that $V = 2958/\lambda_0$, where λ_0 is measured in nanometers. Substituting in Eq. (35), we get

$$D_w = -\frac{1.447 \times 10^4}{\lambda_0} \left[0.080 + 0.549 \left(2.834 - \frac{2958}{\lambda_0} \right)^2 \right] \text{ ps/km.nm}$$

Elementary calculations show that at $\lambda_0 \approx 1300 \text{ nm}$, $D_w = -2.8 \text{ ps/km.nm}$. The variations of D_m , D_w and D_{tot} with λ_0 are shown in Fig. 24.18; the variation of D_m is calculated by using Eq. (31) and Table 8.1. The total dispersion passes through zero around $\lambda_0 \approx 1300 \text{ nm}$ which is known as *zero total dispersion wavelength* and represents an extremely important parameter. At $\lambda_0 \approx 1550 \text{ nm}$, the same fiber will have $D_w = -5.4 \text{ ps/km.nm}$ and $D_m \approx +20 \text{ ps/km.nm}$ giving $D_{\text{tot}} \approx +15 \text{ ps/km.nm}$.

Example 24.18 We next consider the fiber discussed in Example 24.7 for which $n_1 = 1.46$, $\Delta = 0.0075$ and $a = 2.3 \mu\text{m}$, so that $V = 2584/\lambda_0$, where, once again, λ_0 is measured in nanometers. Substituting in Eq. (34) we get

$$D_w = -\frac{3.62 \times 10^4}{\lambda_0} \left[0.080 + 0.549 \left(2.834 - \frac{2584}{\lambda_0} \right)^2 \right] \text{ ps/km.nm}$$

Thus at $\lambda_0 \approx 1550 \text{ nm}$,

$$D_w \approx -19.3 \text{ ps/km.nm}$$

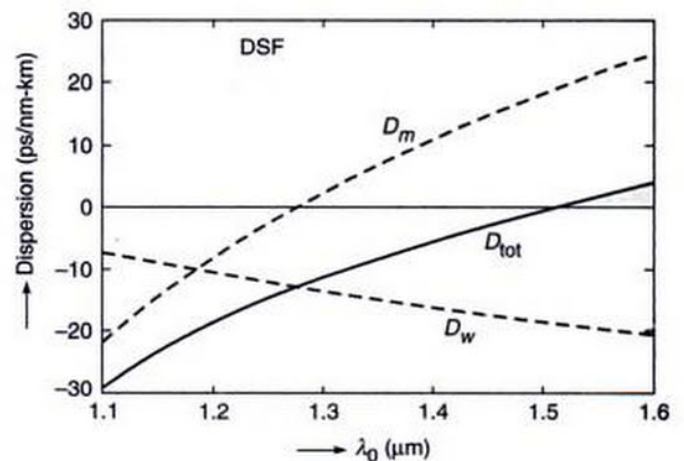


Fig. 24.19 The wavelength dependence of D_m , D_w and D_{tot} for a typical dispersion shifted fiber (DSF) with parameters as given in Example 24.7. The zero dispersion wavelength is around 1550 nm .

*See, e.g., Chapter 10 of Ref. 10.

On the other hand, the material dispersion at this wavelength is given by [see Table 8.1]

$$D_m = +20 \text{ ps/km.nm}$$

We therefore see that the two expressions are of opposite sign and almost cancel each other. Physically, because of waveguide dispersion, longer wavelengths travel slower than shorter wavelengths and because of material dispersion, longer wavelengths travel faster than shorter wavelengths—and the two effects compensate each other resulting in zero total dispersion around 1550 nm. The corresponding variation of D_m , D_w and D_{tot} with wavelength is shown in Fig. 24.19. As can be seen from the figure, we have been able to shift the zero dispersion wavelength by changing the fiber parameters; these are known as the *dispersion shifted fibers*.

24.13 DISPERSION COMPENSATING FIBERS

In many countries there already exist millions of kilometres of conventional single-mode fibers (of the type discussed in Examples 24.6 and 24.17) in the underground ducts operating at 1310 nm; and as mentioned in Example 24.17, these fibers have very low dispersion at the operating wavelength. One can significantly increase the transmission capacity of these systems, by operating these fibers at 1550 nm (where the loss is extremely small) and we can have the added advantage of using EDFA (Erbium Doped Fiber Amplifiers) for optical amplification in this wavelength range (see Sec. 23.1.3). However, if we operate the conventional single-mode fibers at 1550 nm, we will have a significant residual dispersion; and as discussed in Example 24.17, this residual dispersion would be about +15 ps/km.nm. Such a large dispersion would result in significant decrease in the information carrying capacity of the communication system. On the other hand, replacing the existing conventional single-mode fibers by dispersion shifted fibers (DSF's) would involve huge costs. As such, in recent years there has been considerable amount of work in upgradation of the installed 1310 nm optimized optical fiber links for operation at 1550 nm. This is achieved by developing fibers with very large negative dispersion coefficients, a few hundred meters to a kilometre of which can be used to compensate for dispersion over tens of kilometers of the fiber in the link.

In Example 24.18 we have seen that by changing the refractive index profile, we can alter the waveguide dispersion and hence the total dispersion. Indeed, it is possible to have specially designed fibers whose dispersion coefficient (D_{tot}) is large and negative at 1550 nm. A refractive index profile, which is characterized by $D_{\text{tot}} \approx -1,800 \text{ ps/km.nm}$ at 1550 nm, is shown in Fig. 24.20 [Ref. 12]. These types of fibers are known as dispersion compensating fibers

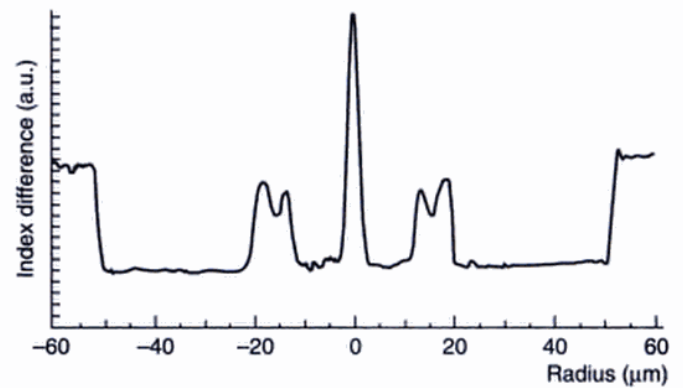


Fig. 24.20 The refractive index profile of a dispersion compensating fiber (DCF) characterized by $D_{\text{tot}} \approx -1800 \text{ ps/km.nm}$ at 1,550 nm [Adapted from Ref. 12].

(DCF's). A short length of DCF can be used in conjunction with the 1310 nm optimized fiber link so as to have small total dispersion value at the end of the link (see Fig. 24.21).

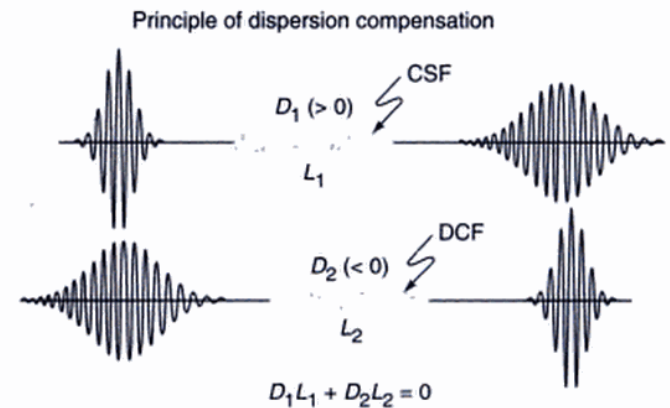


Fig. 24.21 A short length (L_2) of a DCF can be used in conjunction with the conventional single-mode fiber (CSF) so as to have small dispersion value at the end of the link.

In order to understand this phenomenon, we have plotted in Fig. 24.22 (as a solid curve) a typical variation of v_g with wavelength for a conventional single-mode fiber (CSF) with zero dispersion around 1300 nm wavelength (with parameters as given in Example 24.17). As can be seen from the figure, v_g attains a maximum value at the zero dispersion wavelength and on either side it monotonically decreases with wavelength. Thus, if the central wavelength of the pulse is around 1550 nm then the red components of the pulse (i.e., longer wavelengths) will travel slower than the blue components (i.e., smaller wavelengths) of the pulse. Because of this the pulse will get broadened. Now, after propagating through a CSF for a certain length L_1 , the pulse is allowed to propagate through a length L_2 of the DCF

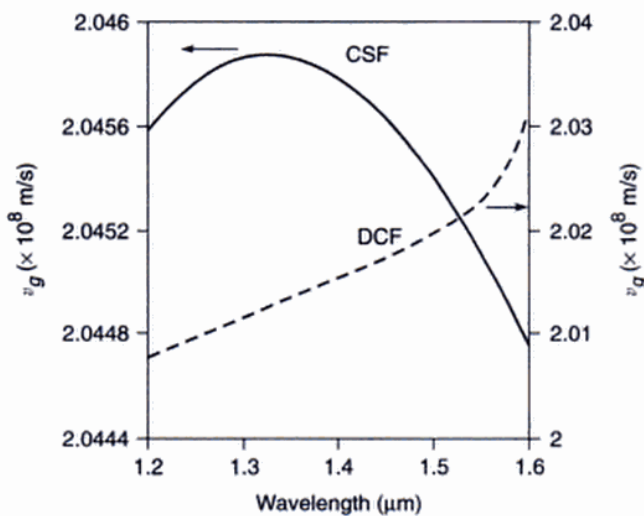


Fig. 24.22 The wavelength variation of group velocity for a typical dispersion compensating fiber and a typical conventional single-mode fiber.

in which the group velocity v_g varies as shown by the dashed curve in Fig. 24.19. The red components (i.e., longer wavelengths) will now travel faster than the blue components and the pulse will tend to reshape itself into its original form. Indeed if the lengths of the two fibers (L_1 and L_2) are such that

$$D_1 L_1 + D_2 L_2 = 0$$

then the pulse emanating from the second fiber will be almost identical to the pulse entering the first fiber as shown in Fig. 24.19 (see also Example 8.6).

24.14 FIBER-OPTIC SENSORS

Although the most important application of optical fibers is in the field of transmission of information, optical fibers capable of sensing various physical parameters and generating information are finding widespread use as fiber-optic sensors. The use of optical fibers for such applications offers the same advantages as in the field of communication viz., lower cost, smaller size, more accurate, greater flexibility and more reliable. As compared to conventional electrical sensors, such fiber-optic sensors are immune to external electromagnetic interference, and can also be used in hazardous and explosive environments. A very important attribute of fiber-optic sensors is the possibility of having distributed or quasi distributed sensing geometries which would otherwise be too expensive or complicated using conventional sensors. Using fiber-optic sensors it is possible to measure pressure, temperature, electric current, rotation, strain, chemical and biological parameters, etc., with

greater precision and speed. These advantages are leading to increased integration of such sensors into civil structures such as bridges and tunnels, process industries, medical instruments, aircrafts, missiles and even cars.

Fiber-optic sensors can be broadly classified into two categories: extrinsic and intrinsic. In the case of extrinsic sensors, the optical fiber simply acts as a device to transmit and collect light from a sensing element, which is external to the fiber. The sensing element responds to the external perturbation and the change in the characteristics of the sensing element is transmitted by the return fiber for analysis. The optical fiber here plays no role other than that of transmitting the light beam. On the other hand, in the case of intrinsic sensors, the physical parameter to be sensed directly alters the properties of the optical fiber, which in turn leads to changes in a characteristic such as intensity, polarisation, phase, etc., of the light beam propagating in the fiber.

There are a large variety of fiber-optic sensors that have been demonstrated in the laboratory and some are already being installed in real systems. In the following, we will discuss some simple extrinsic fiber-optic sensors.

Figure 24.23 shows a very simple sensor based on the fact that transmission through a fiber joint depends on the alignment of the fiber cores. Light coupled into a multi-mode optical fiber couples across a joint into another fiber, which is detected by a photodetector. Any deviation of the fiber pair from perfect alignment is immediately sensed by the detector. A misalignment of magnitude equal to the core diameter of the fiber results in zero transmission. The first 20 % of transverse displacement gives an approximately linear output. Thus for a 50 μm core diameter fiber, approximately 10 μm misalignment will be linear. The sensitivity will of course become better with decrease in core diameter but at the same time, the range of displacements will also reduce.

The misalignment between the fibers could be caused by various physical parameters such as acoustic waves, pressure, etc. Thus if one of the probe fibers has a short free length while the other has a longer length, then acoustic waves impinging on the sensor would set into vibration the fibers which would result in a modulation of the transmitted light intensity leading to an acoustic sensor. Using such an arrangement, deep sea noise levels in the frequency range of 100 Hz to 1 kHz and transverse displacements of a few tenths of nanometres have been measured.¹³ Using the

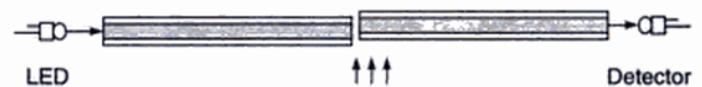


Fig. 24.23 A change in the transverse alignment between two fibers changes the coupling and hence the power falling on the detector.

same principle, any physical parameter leading to a relative displacement of the fiber cores can be sensed using this geometry.

Another very interesting sensor is the liquid level sensor shown in Fig. 24.24. Light propagating down an optical fiber is total internally reflected from a small glass prism and couples back to the return fiber. As long as the external medium is air, the angle of incidence inside the prism is greater than the critical angle and hence light suffers total internal reflection. As soon as the prism comes in contact with a liquid, the critical angle at the prism–liquid interface reduces and the light gets transmitted into the liquid resulting in a loss of signal. By a proper choice of prism material, such a sensor can be used for sensing levels of various liquids such as water, gasoline, acids, oils, etc.

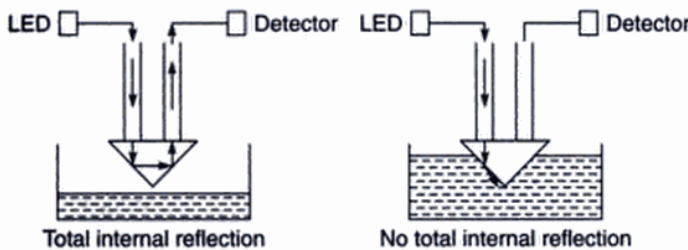


Fig. 24.24 A liquid level sensor based on changes in the critical angle due to liquid level moving up.

SUMMARY

- A step index optical fiber consists of a (cylindrical) central dielectric core cladded by a material of slightly lower refractive index. The corresponding refractive index distribution (in the transverse direction) is given by

$$\begin{aligned} n &= n_1 & 0 < r < a \\ &= n_2 & r > a \end{aligned}$$

where n_1 and n_2 ($n_2 < n_1$) represent the refractive indices of core and cladding respectively and a represents the radius of the core. Light guidance takes place through the phenomenon of total internal reflection at the core-cladding interface.

- A bundle of optical fibers is extensively used in medical endoscopy.
- The numerical aperture (NA) of the fiber, which is a measure of the light gathering power of the fiber, is given by $NA = \sqrt{n_1^2 - n_2^2}$.
- The attenuation of an optical beam is usually measured in decibels (dB). If an input power P_1

results in an output power P_2 , then the loss in decibels

$$\text{is given by } \alpha = 10 \log_{10} \left(\frac{P_1}{P_2} \right)$$

- In commercially available silica fibers, the loss is about 0.5 dB/km at $\lambda_0 \approx 1.3 \mu\text{m}$; the lowest loss (~ 0.3 dB/km) occurs at $\lambda_0 \approx 1.55 \mu\text{m}$. Most fiber optic communication systems operate around the above two wavelengths. EDFAs (Erbium Doped Fiber Amplifiers) are extensively used to amplify signals in the wavelength region 1.53 – 1.57 μm .
- A mode in an optical waveguide is a transverse field distribution which propagates along the fiber without any change in its field distribution except for a change in phase; mathematically it is defined by the equation

$$\Psi(x, y, z, t) = \psi(x, y)e^{i(\beta z - \omega t)}$$

where $\psi(x, y)$ represents the transverse field profile and β represents the propagation constant and the axis of the fiber is along the z -axis. The inverse of the group velocity of the mode is given by:

$$\frac{1}{v_g} = \frac{d\beta}{d\omega}$$

- The dimensionless waveguide parameter V is defined as $V = \frac{2\pi}{\lambda_0} a \sqrt{n_1^2 - n_2^2}$, where λ_0 is the free space wavelength. For a step index fiber, if $V < 2.4045$ the fiber supports only one mode and the fiber is said to be a single mode fiber. The value of λ_0 for which a given step index fiber has $V = 2.4045$ is known as the cut off wavelength and is denoted by λ_c . For $\lambda_0 > \lambda_c$, we have a single mode fiber. On the other hand, for $V \geq 10$, the number of modes is approximately given by $\frac{1}{2} V^2$ and the fiber is said to be a multimoded fiber.
- In multimode fibers, different rays take different times to propagate through a given length of the fiber; this leads to what is known as *intermodal dispersion* because it arises due to different modes traveling with different group velocities. The intermodal dispersion is given by

$$\begin{aligned} \Delta\tau_i &= \frac{n_1 L}{c} \left(\frac{n_1 - n_2}{n_2} \right) \quad \text{for a step index fiber} \\ &= \frac{n_2 L}{c} \left(\frac{n_1 - n_2}{n_2} \right)^2 \quad \text{for a parabolic index fiber} \end{aligned}$$

where L is the length of the fiber and c ($\approx 3 \times 10^8$ m/s) is the speed of light in free space. Typically, $(n_1 - n_2)/n_2 \approx 0.01$ and $\Delta\tau_i \approx 50$ ns/km for a step index fiber and ≈ 0.25 ns/km for a parabolic index fiber. It

10. A. Ghatak and K. Thyagarajan, *Introduction to Fiber Optics*, Cambridge University Press, Cambridge, 1998.
11. D. Marcuse, 'Loss analysis of single-mode fiber splices', *Bell Syst. Tech. J.*, Vol. 56, 703, 1977.
12. J. L. Auguste, R. Jindal, J. M. Blondy, M. Clapeau, J. Marcou, B. Dussardier, G. Monnom, D. B. Ostrowsky, B. P. Pal and K. Thyagarajan, '-1800 ps/(nm.km) chromatic dispersion at 1.55 μm in dual concentric core fiber', *Electronics Letters*, Vol. 36, 1689, 2000.
13. W. B. Spillman, and R. L. Gravel, 'Moving fiber-optic hydrophone', *Optics letters*, Vol. 5, 30-33, 1980.
14. B.P.Pal (Ed.), *Fundamental of Fiber Optics in Telecommunications and Sensor Systems*, Wiley Eastern Ltd, New Delhi (1992).

Introduction to Speckle Metrology*

The granular pattern produced by a laser beam encountered by a diffusing object, often called laser speckles, was famous as a major source of noise in laser optics till 1970. Now it has revolutionised the whole field of metrology; particularly because, the analysis is often simpler than that in holography. Due to the high degree of coherence of laser light, the light reaching at any point from the entire surface interferes forming an intensity variation. This variation combined by the resolution of the recording system (such as eye, camera, photographic film, and such like) finally gives the random granular structure. Figure 25.1 shows a typical magnified speckle pattern. Thus even an uniformly illumi-

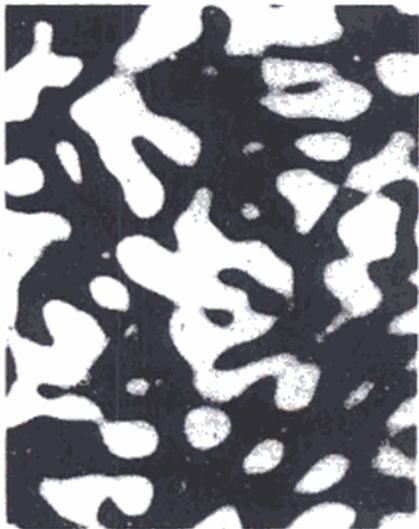


Fig. 25.1 A typical magnified speckle pattern on the image of a rough metal surface illuminated uniformly by He-Ne laser ($\lambda = 6328 \text{ \AA}$). The minimum resolvable speckle diameter was about $0.8 \text{ }\mu\text{m}$.

nated surface appears with a noisy background. One can easily observe fairly big speckles by observing the surface through a pinhole. This is due to the loss of resolution created by the small aperture.

The contrast of the speckles is zero for a perfectly reflecting surface and it increases as the surface roughness increases. However, even when the roughness is well within the wavelength of the light, the contrast becomes unity and remains so for any higher roughness. Thus, the contrast measurement has proved to be a powerful tool for the nondestructive testing of small surface roughnesses within the light wavelength.

The other important applications include displacement or motion analysis and relevant nondestructive testing applications. To illustrate, we are discussing here the simple example of the well-known double-speckle laser-speckle photography for lateral (in plane) displacement analysis. Historically, the method involving a single illuminating beam is called photography whereas with two beams it is called interferometry although both are interferometric methods.

For this, the surface to be studied is illuminated by a divergent laser beam (Fig. 25.2). Laser speckles are formed on the camera focussed on the film plane. The smallest speckle size S_0 on the object plane will thus be governed by the lens resolution, given by

$$S_0 \approx 1.2 \lambda F(m + 1) \quad (1)$$

where λ is the wavelength, F the aperture ratio and m the demagnification (object/image ratio). Obviously we assume that the film resolution is capable of recording the pattern.

If a double exposure photograph with displaced object is recorded on the same film we get two different structures. However, it is known that for a lateral displacement of the

*The entire chapter has been very kindly written by Professor C.S. Vikram of University of Alabama at Huntsville, Alabama, USA.

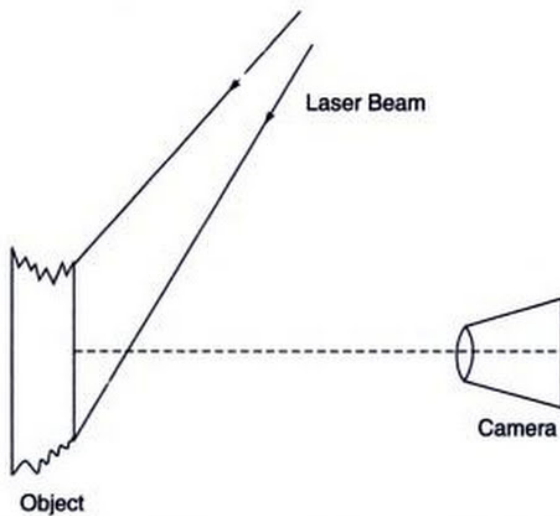


Fig. 25.2 Schematic arrangement for single beam laser speckle photography.

object, the nature of the random structure on the image plane remains unchanged and therefore we get two photographs of the same structure but laterally displaced.

The analysis of the transparency thus formed will give the displacement of the object. There are two methods for the analysis. The first is to analyse the entire surface by Fourier filtering. The simplest way to achieve this is to put the transparency in a converging beam and see or record the photograph through an aperture in the focal plane (Fig. 25.3).

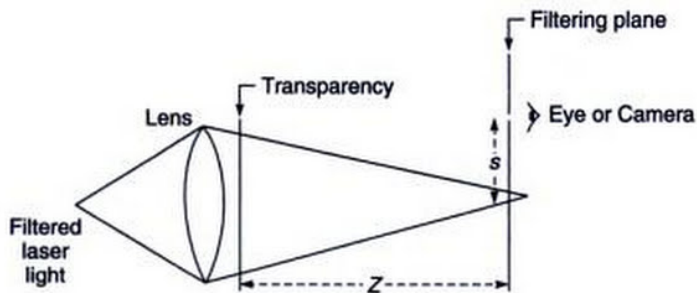


Fig. 25.3 Fourier filtering (or full view) approach of the analysis of speckle photographs.

Fringes are seen whose frequency and orientation give the displacement. These fringes are the loci of the displacements in the direction along the line joining the centre of the filtering plane and the aperture. The displacement u corresponding to a fringe is

$$u = \frac{\lambda z}{ms} \quad (2)$$

where z is the distance between the photographic transparency and the filtering plane, and s is the distance of the filtering aperture from the optical axis. The advantages

of this method are that the entire surface can be studied simultaneously and that the fringe frequency can be varied by the distance s of the aperture from the centre.

The other method of the analysis is a point-wise scanning of the photograph (Fig. 25.4). This is very popular due to its simplicity. A point of the transparency is simply illuminated by a laser beam. A diffraction halo with a set of Young's fringes is observed at a distance on the screen.

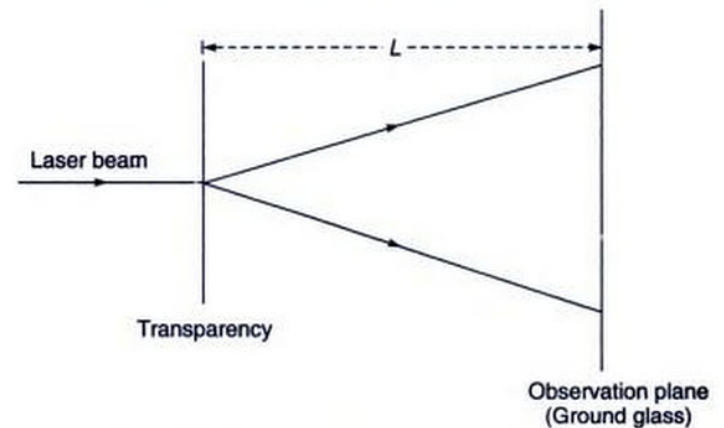


Fig. 25.4 Young's fringes (point-wise scanning) method of the analysis of speckle photographs.

The fringes are the measure of the object displacement D given by

$$D = \frac{\lambda L}{md}$$

where d is the fringe spacing and L is the distance from the screen to the transparency. The orientation gives the direction of the displacement (both are perpendicular to each other). A typical set of Young's fringes is given in Fig. 25.5.



Fig. 25.5 A typical Young's fringe pattern. There is a diffraction halo containing fringes relating to lateral displacement.

There is yet another method. This involves illuminating the object by two beams, particularly to reduce the measurement range and to observe the changes in real time. There are methods to process the image electronically (no

photographic recording) using a TV vidicon under the subject *Electronic Speckle Pattern Interferometry*. Attempts have also been made to eliminate the need for wet chemical processing and still use the conventional manner of analysis. These are replacing the usual photographic film by instant films, thermoplastic photographic materials, liquid-crystal light valves, BSO crystals, photographic diffusers, real time heterodyne approach, etc. There is a method of magnifying the speckles using lenses and a TV camera-monitor system to observe these movements

directly. However, a detailed discussion of these techniques is beyond the scope of this text.

SUMMARY

- The granular pattern produced by a laser beam encountered by a diffusing object is known as laser speckles. Applications of laser speckles include displacement or motion analysis and also in nondestructive testing.

REFERENCES AND SUGGESTED READINGS

1. J.C. Dainty, (Ed.), *Laser Speckle and Related Phenomena*, Springer-Verlag, Berlin and New York, 1975.
2. R.K. Erf, (Ed.), *Speckle Metrology*, Academic Press, New York, 1978.
3. M. Francon, *Laser Speckle and Applications in Optics*, Translated by H. H. Arsenault, Academic Press, New York, 1979.

Since

$$\Gamma(1) = \int_0^{+\infty} e^{-x} dx = 1$$

we obtain

$$\Gamma(n+1) = n! ; n = 0, 1, 2, \dots$$

Further since $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$, we obtain

$$\left. \begin{aligned} \Gamma\left(\frac{3}{2}\right) &= \frac{1}{2} \Gamma\left(\frac{1}{2}\right) = \frac{1}{2} \sqrt{\pi} \\ \Gamma\left(\frac{5}{2}\right) &= \frac{3}{2} \Gamma\left(\frac{3}{2}\right) = \frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi} \\ \Gamma\left(\frac{7}{2}\right) &= \frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \sqrt{\pi} \end{aligned} \right\} \quad (9)$$

etc. Finally for $n = 0, 1, 2, \dots$

$$\int_{-\infty}^{+\infty} x^{2n} e^{-x^2} dx = \Gamma\left(n + \frac{1}{2}\right) = \frac{1.3.5 \dots (2n-1) \sqrt{\pi}}{2^n} \quad (10)$$

and

$$\int_{-\infty}^{+\infty} x^{2n+1} e^{-x^2} dx = 0 \quad (11)$$

Appendix B

Diffraction of a Gaussian Beam

If the amplitude and phase distribution on the plane $z = 0$ is given by $A(\xi, \eta)$ then the diffraction pattern is given by [see Eq. (35) of Chapter 17]

$$u(x, y, z) = -\frac{i}{\lambda z} \exp(ikz) \iint A(\xi, \eta) \exp \left\{ +\frac{ik}{2z} [(x-\xi)^2 + (y-\eta)^2] \right\} d\xi d\eta \quad (1)$$

We consider a Gaussian beam propagating along the z -direction whose amplitude distribution on the plane $z = 0$ is given by

$$A(\xi, \eta) = a \exp \left[-\frac{\xi^2 + \eta^2}{w_0^2} \right] \quad (2)$$

implying that the phase front is plane at $z = 0$. Thus at a distance w_0 from the z -axis, the amplitude falls by a factor $1/e$ (i.e., the intensity reduces by a factor $1/e^2$). This quantity w_0 is called the *spot size* of the beam. Substituting Eq. (2) in Eq. (1), we obtain

$$u(x, y, z) = -\frac{ia}{\lambda z} e^{ikz} \int_{-\infty}^{\infty} \exp \left[\frac{ik}{2z} (x-\xi)^2 - \frac{\xi^2}{w_0^2} \right] d\xi \\ \times \int_{-\infty}^{\infty} \exp \left[\frac{ik}{2z} (y-\eta)^2 - \frac{\eta^2}{w_0^2} \right] d\eta$$

or

$$u(x, y, z) = -\frac{iae^{ikz}}{\lambda z} e^{\frac{ik}{2z}(x^2+y^2)} \int_{-\infty}^{\infty} e^{-\alpha\xi^2 + \beta_1\xi} d\xi \\ \times \int_{-\infty}^{\infty} e^{-\alpha\eta^2 + \beta_2\eta} d\eta \quad (3)$$

where

$$\alpha = \frac{1}{w_0^2} - \frac{ik}{2z} = -\frac{ik}{2z} (1 + i\gamma) \quad (4)$$

$$\gamma = \frac{\lambda z}{\pi w_0^2} \quad (5)$$

$$\beta_1 = -\frac{ikx}{z}, \quad \beta_2 = -\frac{iky}{z}$$

If we now use the integral

$$\int_{-\infty}^{\infty} e^{-\alpha x^2 + \beta x} dx = \sqrt{\frac{\pi}{\alpha}} \exp \left[\frac{\beta^2}{4\alpha} \right] \quad (6)$$

we would get

$$u(x, y, z) = \frac{a}{(1+i\gamma)} \exp \left[-\frac{x^2 + y^2}{w^2(z)} \right] e^{i\Phi} \quad (7)$$

where

$$w(z) = w_0 [1 + \gamma^2]^{1/2} = w_0 \left[1 + \frac{\lambda^2 z^2}{\pi^2 w_0^4} \right]^{1/2} \quad (8)$$

and

$$\Phi = kz + \frac{k}{2z} (x^2 + y^2) - \frac{k(x^2 + y^2)}{2z(1+\gamma^2)} \\ = kz + \frac{k}{2R(z)} (x^2 + y^2) \quad (9)$$

where

$$R(z) = z \left[1 + \frac{1}{\gamma^2} \right] = z \left[1 + \frac{\pi^2 w_0^4}{\lambda^2 z^2} \right] \quad (10)$$

Index

A

Aberrations [5.1](#)
Achromatic doublet [5.2](#)
Active medium 23.5
Airy pattern 16.9
Anisotropic media 19.24
 ray paths in [2.18](#)
 wave analysis of 19.25
Aplanatic points [3.8](#), [3.11](#)
Astigmatism [5.11](#)
Attenuation in optical fibers 24.11

B

Birefringence – see Double refraction
Birefringent fibers 19.20
Bit rates 24.17
Blue shifted pulse [8.12](#)
Bragg's law 13.9, 16.27
Brewster's law 19.7, 21.4

C

Cartesian oval [3.10](#)
Cavity lifetime 23.18
Characteristics of lasers 23.3
Chirping [8.10](#)
Chromatic aberration [5.1](#)
Circular aperture, diffraction by 16.8, 17.3, 17.6, 17.20
Circularly polarized light 19.11
Coherence length [15.1](#)
Coherence time [15.1](#)
 and linewidth [15.10](#)
Coherence 12.5, [15.1](#)
Coherent bundle 24.9
Collisional broadening 23.21
Colours of thin films 13.16
Coma [5.9](#)
Complex degree of coherence [15.11](#)
Complex representation [11.9](#)

Compton effect 22.6
Compton scattering, kinematics of 22.8
Conducting medium, wave equation in 20.10
Constitutive relations 20.1
Continuity conditions 20.11
Cornu's spiral 17.13
Corpuscular model [1.3](#)
Cosine law 13.3
Current sensor 19.32
Curvature of field [5.11](#)

D

Damped simple harmonic motion 6.7
Debye-Scherrer rings 16.29
Dielectric film, reflectivity of 13.6, 21.14
Diffraction by a circular aperture 16.8, 17.3, 17.20
Diffraction by a rectangular aperture 17.19
Diffraction by a single slit 16.4
 transition to the Fraunhofer region 17.16
Diffraction by a straight edge 17.11
Diffraction by an opaque disc 17.4, 17.7
Diffraction grating 16.23
 oblique incidence on 16.26
 resolving power of 16.24
Diffraction of a Gaussian beam 17.8, [B.1](#)
Diffraction of water waves [10.3](#)
Dipole radiation pattern 20.7
Directionality of laser beams 16.10
Dispersive medium [8.8](#)
Dispersion compensating fibers 24.20
Distortion [5.12](#)
Doppler broadening 23.20
Double refraction 19.3, 19.12
Double-slit Fraunhofer diffraction pattern 16.17
Down chirped pulse [8.12](#)

E

EDFA 23.6

gain flattening of 23.8
 gain spectrum of 23.8
 Einstein coefficients 23.15
 Electromagnetic waves 20.1
 energy density of 20.6
 reflection and refraction of 21.1
 Ellipsoidal reflector 2.7
 Elliptic core fiber 19.20
 Elliptically polarized light 19.11
 Endoscope 24.10
 Energy density of an electromagnetic wave 20.6
 Erbium doped fiber amplifier – see EDFA
 Evanescent wave 21.8
 Extraordinary ray 19.29
 Extraordinary wave 19.28

F

Fabry-Perot cavity, modes of 14.5
 Fabry-Perot [etalon](#), 14.3
 Fabry-Perot interferometer 14.5
 resolving power of 14.7
 Faraday rotation 19.32, 19.34
 Fermat's principle 2.3
 Fiber Bragg gratings 13.10
 Fiber endoscope 24.10
 Fiber laser 23.8
 Fiber optic sensors 24.21
 current sensor 19.32
 Fiber optics 24.3
 Focusing of laser beams 16.12
 Forced vibrations 6.9, 7.5
 Fourier integral theorem 7.7
 Fourier integral 7.6
 Fourier series 7.1
 Fourier transform spectroscopy 15.13
 Fourier transform 7.7, 15.13
 Fraunhofer diffraction 16.3
 Frequency spectrum 7.7
 Frequency, concept of 9.3
 Fresnel diffraction 17.1
 Fresnel biprism 12.15
 Fresnel half-period zones 17.2
 Fresnel integrals 17.12
 Fresnel's two mirror arrangement 12.14
 Fundamental mode, spot size of 24.13

G

Gain spectrum 23.8
 Gamma functions [A.1](#)
 Gaussian beam, diffraction of 17.8, [B.1](#)
 Gaussian formula 3.3
 Gaussian function, Fourier transform of 7.7
 Gaussian spot size 24.14
 Graded index atmosphere 2.12

Graham Bell's experiment 24.5
 Graphical method 11.7
 Grating, see Diffraction grating
 Grazing incidence 21.5
 GRIN lenses 2.14
 Group velocity of a wave packet [8.5](#)
 Group velocity 8.1

H

Half wave plates 19.16
 Helium Neon laser – see He-Ne laser
 He-Ne laser 23.10, 23.20
 Hertz's experiment 1.6
 High reflectivity by thin film deposition 13.7
 Historical remarks on fiber optics 24.4
 Hologram, formation of 18.4
 Holography 18.1
 Holography, applications of 18.7
 Huygens' eyepiece 5.9
 Huygens' principle 10.1
 Huygens-Fresnel principle 17.1

I

Interference by a parallel film 13.1
 Interference experiments, understanding of [1.11](#)
 Interference filters 14.10
 Interference of light, by division of amplitude 13.1
 by division of wavefront 12.1
 Interference of polarized light 19.16
 Interference with white light 12.15
 Intermodal dispersion 24.15
 Ionosphere 2.16, 6.18
 reflections from 2.16
 Ives experiment 11.6

J

Jones calculus 19.30

L

Laser beam, monochromaticity of 23.23
 directionality of 16.10
 focusing of 16.12
 Laser speckles 25.2
 Lasers 23.3
 Lasing action 23.7
 Lateral magnification 3.7
 Laue spots 16.30
 Laws of [reflection](#), 2.4
 Laws of refraction 2.4
 Lens formula by Huygens' principle 10.8
 Limit of resolution 16.14
 Line shape function 23.20
 Linearly polarized wave 19.4, 19.11
 Linewidth 15.3, 15.10
 Lloyd's mirror arrangement 12.17

Longitudinal sound waves in a gas 9.9
 Longitudinal sound waves in a solid 9.7
 Looming 2.11
 Lummer-Gehrcke plate 14.9

M

Malus' law 19.8
 Material dispersion 8.4, 24.17
 Matrix method in paraxial optics 4.1
 Matter waves 1.9
 Maximum bit rates 24.17
 Maxwell's equations 20.1
 physical significance of 20.12
 Meter, standardization of 13.22
 Michelson interferometer 13.21, 15.2
 Michelson stellar interferometer 15.6
 Microscope, resolving power of 16.17
 Miller indices 16.28
 Mirage 2.9
 Modes of resonator 23.12
 Moiré fringes 12.12, 15.8
 Momentum of a photon, 20.9, 22.6
 Momentum of an electromagnetic wave 20.9
 Monochromatic aberrations 5.4
 Multi-mode fibers 24.12
 Multiple beam interferometry 14.1
 Multiple slit Fraunhofer diffraction pattern 16.20

N

Natural broadening 23.21
 Newton formula 3.7
 Newton's rings 13.16
 Nodal curves 12.4
 Nodal planes 4.8
 Non-dispersive medium 8.6
 Non-reflecting films 13.4
 N-slit Fraunhofer diffraction pattern 16.20
 Numerical aperture 24.10

O

Oblique incidence on a grating 16.26
 Oil immersion objective 3.9
 One dimensional wave equation 9.6
 solution of 9.10
 Opaque disc, diffraction by 17.4, 17.7
 Optic axis 19.13
 Optical activity 19.20
 Optical theory of 19.33
 amplification 23.6, 23.15
 Optical beats 15.7
 Optical fiber 24.3
 Optical resonator 23.5, 23.7, 23.11
 Ordinary ray 19.29
 Ordinary wave 19.28

Origin of refractive index 6.11
 Oscillating dipole 20.7

P

Parabolic index fibers 24.16
 Parabolic index media 2.13
 Paraboloidal reflector 2.6
 Particle nature of radiation 1.3, 1.7, 22.1
 Periodic structure, reflection by 13.8
 Phase change on reflection 12.17, 21.4
 Phase velocity 8.1
 Photoelectric effect 22.4
 Photon mass 22.9
 Photon 1.12, 22.5
 momentum of 20.9, 22.6
 polarization of 1.12
 Photophone 24.5
 Photosensitivity 13.10
 Plane waves 9.12
 in a dielectric 20.2
 propagation in anisotropic media, 19.24
 Poisson spot 17.4
 Polarization by double refraction 19.7
 Polarization by reflection 19.7
 Polarization by scattering 19.8
 Polarization of light 19.3
 Polarized light, analysis of 19.19
 production of 19.6
 Polaroid 19.6
 Population inversion 23.16
 Poynting vector 20.5
 Principal foci 3.5
 Prism film coupling experiment 24.13
 Prism, resolving power of 16.25
 Probabilistic interpretation of matter waves 1.9
 Propagation in dispersive media 8.8
 Pulse dispersion 8.1, 24.14
 Pumping source 23.5

Q

Quarter wave plates 19.16

R

Radiation pressure 20.9
 Raman amplification 23.24
 Raman scattering 23.24
 Ray dispersion 24.15
 Ray equation 2.12
 Ray paths in inhomogeneous medium 2.8
 Ray paths in parabolic index media 2.13
 Rayleigh criterion of resolution 16.15
 Rayleigh scattering 6.15, 24.11
 Rectangular aperture 17.19
 Rectilinear propagation 10.2
 Red shifted pulse 8.12
 Reflection by a conducting medium 21.13

Reflection by a periodic structure 13.8
 Reflection by a spherical surface 3.3
 Reflection by Huygens' principle 10.5
 Reflection of electromagnetic waves 21.1
 Reflectivity of a dielectric film 13.6, 21.14
 Reflectivity of a film 13.6
 Refraction by a spherical surface 3.1
 Refraction by Huygens' principle 10.4, 10.6
 Refraction of electromagnetic waves 21.1
 Refractive index, origin of 6.11
 Resolving power of a grating 16.24
 Resolving power of a prism 16.25
 Resolving power of microscope 16.17
 Resonance 6.10
 Resonator modes 23.12
 Ripple tank interference pattern 12.3
 Rochon prism 19.24
 Rods and cones 24.9
 Ruby laser 23.9, 23.22
 typical parameters of 23.22

S

Sagittal plane 5.11
 Scanning Fabry-Perot interferometer 14.7
 Self-focusing phenomenon 16.30
 Shadow region 2.11
 Sign convention 3.2
 Simple harmonic motion 6.3
 Sine condition 3.12
 Single mode fibers 24.12
 Single slit diffraction experiment 1.8, 16.4
 and uncertainty principle 1.8
 Sinusoidal waves 9.3
 superposition of 11.6
 Spatial coherence 15.4
 Spatial frequency filtering 16.32
 Speckle metrology 25.1
 Spherical aberration 5.4
 Spherical refracting surface, imaging by 4.4
 Spiking in ruby laser 23.10
 Splice loss 24.14
 Spontaneous emission 23.4, 23.15
 Spot size of the fundamental mode 24.13
 Square law media 2.13
 Stability diagram 23.14
 Standardization of meter 13.22
 Standing electromagnetic waves 11.4
 Standing waves 11.4, 11.5
 Stationary light waves 11.6
 Stationary waves 11.4, 11.5
 Stellar interferometer 15.6
 Step index fibers 24.8
 pulse dispersion of 24.15

Stimulated absorption 23.4, 23.15
 Stimulated emission 23.4, 23.15
 Stokes relations 12.18
 Straight edge diffraction pattern 17.11
 String, transverse vibrations of 7.3, 9.6, 11.5
 Superposition of two sinusoidal waves 11.6
 Superposition of waves 11.5, 19.9
 Surface waves 21.8

T

Thin films, colors of 13.16
 Thin lens 3.4
 Three-dimensional wave equation 20.4
 solution of 9.11
 Threshold condition 23.19
 Time-energy uncertainty relation 1.14
 Total internal reflection 21.5, 24.6
 by Huygens' principle 10.5
 Transit time calculations in parabolic index media 2.15
 Transverse electromagnetic waves 20.2
 Transverse misalignment 24.14
 Transverse vibrations of a plucked string 7.3
 Two beam interference 12.1
 Two-slit Fraunhofer diffraction pattern 16.17

U

Uncertainty relation 1.8, 1.13, 1.14
 Uniaxial crystals 19.13, 19.27, 19.30
 Unit planes 4.7

W

Water jet experiment 24.7
 Wave equation 9.6, 9.10
 in a conducting medium 20.10
 solutions of 9.10, 9.13
 Wave model 1.5
 Wave motion 9.4
 Wave packet, group velocity of 8.5
 Wave propagation 9.1
 in anisotropic media, 19.24
 Waveguide dispersion 24.18
 Wavelength, concept of 9.3
 Waves, superposition of 11.3
 Wiener's experiment 11.6
 Wire grid polarizer 19.6
 Wollaston prism 19.23

X

X-ray diffraction 16.27

Y

Young's arrangement 12.7
 Young's double-hole experiment 12.7, 15.2

Z

Zero material dispersion 8.5
 Zone plate 17.4

3rd
EDITION

OPTICS

From the reviews of the earlier editions:

"...Accordingly need has arisen for suitable new textbooks on optics. Ajoy Ghatak's OPTICS is an admirable effort in this direction...The treatment throughout the book is both thorough and authoritative..." G C Trigonayat, *Journal of Physics Education*.

"The writing is clear and free flowing...Overall it has been a pleasure to read the book..." E S R Gopal, *University News*.

"The fascinating tale of light
Put marvelously in black and white
Will give students and teachers delight."
Satyanarayan Acharya, *Physics Education*.

This comprehensive and thoroughly revised edition would meet the requirements of undergraduate students of science and engineering. Researchers involved in general areas of optics and laser would find this book immensely useful.



Ajoy Ghatak is currently Emeritus Professor of Physics at IIT Delhi. He obtained his MSc from Delhi University and PhD from Cornell University. He has published over 170 research papers and several books. His areas of interest are fiber optics and quantum mechanics.

Professor Ghatak is a recipient of several national and international awards which include the 2003 Optical Society of America Esther Hoffman Beller award for his outstanding contribution to optical science and engineering education.

The McGraw-Hill Companies

ISBN 0-07-058583-0



9 780070 585836

<http://highered.mcgraw-hill.com/sites/0070585830>

Visit **Tata McGraw-Hill** at:
www.tatamcgrawhill.com